学校编码：10384

学　　号：23320180155680

厦門大學

博　士　学　位　论　文

基于 MAB 技术的无线网络资源分配与优化策略研究

Research on Resource Allocation and Optimization Strategy Based on Multi-Armed Bandit in Wireless Networks

童景文

指导教师姓名：　付 立 群 教 授

专 业　名 称：　通信与信息系统

论文提交日期：　2022　年　4　月

论文答辩日期：　2022　年　5　月

学位授予日期：　2022　年　6　月

2022 年　5　月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果,均在文中以适当方式明确标明,并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（厦门大学付立群教授）课题（组）的研究成果，获得（国家自然科学基金、福建省百人计划科研项目）课题（组）经费或实验室的资助，在（厦门大学未来网络空间 FUNLab）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

本人声明该学位论文不存在剽窃、抄袭等学术不端行为,并愿意承担因学术不端行为所带来的一切后果和法律责任。

声 明 人（签名）：

指导教师（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的涉密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（ √ ）2.不涉密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。涉密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

学位论文答辩委员会名单

主席	王　琳	厦门大学	教授	博士生导师
委员	唐余亮	厦门大学	教授	博士生导师
	程　恩	厦门大学	教授	博士生导师
	袁晓军	电子科技大学	教授	博士生导师
	陈平平	福州大学	教授	博士生导师
秘书	王德清	厦门大学	助理教授	硕士生导师

答辩时间：2022 年 5 月 9 日下午

答辩地点：厦门大学海韵园行政楼 C-501 会议室

摘 要

随着日益复杂的国际形势和严峻的竞争压力，第六代移动通信系统（6th Generation Mobie Networks, 6G）的研发已成为各国竞相发展的一项国家战略。6G将引入更高的频谱、能量、覆盖效率，更强的安全性以及新的使能技术，使得网络变得更加致密化、异构化、智能化，从而给无线网络中的资源分配与优化带来一系列挑战，如高密度异构化网络和高可靠性低延时应用等。为了克服上述挑战，本论文围绕下一代无线通信网络中几种典型应用场景，针对全双工载波侦听多路访问（Full Duplex Carrier Sensing Multiple Access，FD-CSMA）网络空间复用、分布式异构网络资源分配、高密度物联网设备调度、水声通信网络链路自适应四个问题，结合优化理论、博弈理论、马尔科夫决策过程、数据驱动技术，提出一类基于多臂老虎机（Multi-Armed Bandit，MAB）技术的资源分配与优化算法，为实现下一代无线通信网络的愿景提供理论依据和技术支撑。研究成果包括：

（1）针对 FD-CSMA 网络中的空间复用问题，通过联合考虑 FD 链路上的传输功率控制、载波侦听门限调整和对数接入强度自适应，结合优化理论和 MAB 技术，本文提出了一种在随机和对抗环境中均最优的 FD-CSMA 算法。首先，利用分解理论将该空间复用问题分解成 MAC 层的调度问题和物理层的参数选择问题。其次，针对 MAC 层的调度问题，基于拉格朗日乘子法和次梯度下降法，设计了一种最优 FD-CSMA 算法来获得链路的最佳对数接入强度；针对物理层的参数选择问题，将其建模成多个玩家的对抗 MAB 框架，提出了一种在随机和对抗环境中均最优的 MAB 算法来得到链路的最佳发送功率和载波侦听门限。最后，通过交替迭代求解这两个子问题得到每条全双工链路的最佳发送功率、载波侦听门限和对数接入强度，最大化网络的空间复用率。本文理论分析了算法的遗憾上界，且数值结果表明所提算法的网络平均吞吐量比随机选择方法提高了约 43%。

（2）针对分布式异构网络中的资源分配问题，通过考虑物联网设备的智能反射面和扩频因子的联合选择问题，结合博弈理论和 MAB 技术，本文提出了一种完全分布式的 E2Boost（Exploration and Exploitation Boosting）资源分配算法。首先，利用在线学习理论将该联合选择问题建模成两阶段的多玩家 MAB（Multi-Player

MAB，MPMAB）框架。其次，针对第一阶段 MPMAB，基于贪婪算法和非合作博弈方法，提出了一种最优智能反射面分配策略。在此基础上，针对第二阶段 MPMAB，基于汤普森采样算法，提出了一种最优扩频因子分配策略。然后，通过交替迭代求解该两阶段 MPMAB 问题，提出了一种完全分布式的 E2Boost 算法，得到物联网设备的最佳智能反射面和扩频因子，最大化网络中物联网设备的传输速率之和。理论分析了 E2Boost 算法的遗憾上界；仿真结果表明所提算法的性能比已有的分布式分配策略提高了约 M 倍，其中 M 是物联网设备的扩频因子数目。

（3）针对高密度物联网中基于信息新鲜度（Age-of-Information，AoI）的设备调度问题，结合马尔科夫决策过程和 MAB 技术，本文提出了一种基于广义惠特尔索引的调度策略。首先，考虑物联网设备之间的相关性，将该设备调度问题建模成带有相关性的马尔科夫 MAB 框架。其次，基于分解理论，通过为每个物联网设备引入一个中间状态将该高维的马尔科夫 MAB 问题解耦成多个一维的子问题。接着，在理想和非理想信道模型下，通过求解每一个子问题的贝尔曼方程，得到了广义惠特尔索引的解析表达式，进而提出了一种基于广义惠特尔索引的调度策略，最小化网络的平均 AoI。最后，理论分析了所提策略的性能下界；仿真结果表明在高密度网络中所提策略的性能显著优于已有的基于 AoI 的调度策略。

（4）针对水声通信网络中的链路自适应问题，考虑链路的传输频率和速率的联合选择问题，结合问题模型特征、数据驱动技术和 MAB 技术，本文提出了一类收敛速度快、计算复杂度低的联合选择算法。首先，利用在线学习理论将该联合选择问题建模成单个玩家的 MAB 框架。在平稳水声信道下，利用目标函数的二维单峰特性提出了一种基于单峰特性的 MAB 算法；在非平稳水声信道下，利用联合缓变信道追踪和突变信道检测技术提出了一种基于混合非平稳检测和单峰特性的 MAB 算法；在动作空间较大且目标函数不存在单峰特性的情况下，利用动作之间的相关性和数据驱动技术提出了一种迭代边界收缩的 MAB 算法。其次，理论分析了所提算法的遗憾上界，并表明其收敛速度比传统的 MAB 算法提高了约 $\log_2\left(MN/5\right)$ 倍，其中 M 和 N 分别是传输频率和速率的数目；最后，数值结果验证了理论分析的正确性，且表明所提算法比传统 MAB 算法具有较快的收敛速度。

关键词：无线通信网络；资源分配与优化；MAB 技术；优化理论

Abstract

With the increasingly complex international environment and severe competitive pressure, the 6th Generation Mobile Networks (6G) development has become a national strategy for the world around. 6G will introduce higher spectrum/energy/coverage efficiency, more robust security, new enabling technologies, and ubiquitous intelligence, making the network highly dense, heterogeneous, and intelligent. However, 6G will also bring seveal challenges for the resource allocation and optimization problems in the wireless networks, such as high-density heterogeneous networks and reliable and low-latency applications. This thesis focuses on spatial reuse of full-duplex carrier sensing multiple access (FD-CSMA) networks, distributed resource allocation of heterogeneous cellular networks, device scheduling of high-density IoT networks, and link adaptation of underwater acoustic communication networks. To overcome these challenges, we put forth a series of allocation and optimization stategies by combining the communication theory, optimization theory, game theory, data-driven technology, and multi-armed bandit (MAB) technology, dedicated to realize the vision of next-generation wireless communication networks and to provide the theoretical and technical support. The main contributions of this thesis are summarized as follows.

(1) To solve the spatial reuse problem in FD-CSMA network, this thesis proposes a stochastic and adversarial optimal (SAO) FD-CSMA algorithm to allocate the optimal transmit power (TP), carrier sensing threshold (CST), and logarithmic access intensity (LAI) to each FD link by combining with the optimization theory and MAB technolog. First, this spatial reuse problem is decomposed into a link scheduling problem in the MAC layer and a parameter selection problem in the physical layer based on the decomposition method. For the scheduling problem, this thesis designs an LAI adaptive algorithm using the Lagrange multiplier method and the subgradient descent method. For the parameter selection problem, we model it as a multi-player adversarial MAB framework and propose a SAO-based MAB algorithm to find the optimal TP and CST for each FD link. Then, the optimal TP, CST, and LAI can be determined on each FD link by alternately iteratively solving these two sub-problems. Finally, the theoretical result provides an upper regret bound for the proposed algorithm. In addition, the numerical results show that the proposed algorithm can improve the network throughput by 43% compared with the random selection method.

(2) To handle the joint reconfigurable intelligent surface (RIS) and spreading factor (SF) allocation problem in the distributed heterogeneous networks, this thesis proposes an Exploration and Exploitation Boosting (E2Boost) algorithm by combining the non-cooperative game theory and the MAB technology. First, the joint selection problem is modeled as a two-stage Multi-Player MAB (MPMAB) problem using the MAB technique. The first MPMAB problem is to find the best RIS for each IoT device using the epsilon greedy algorithm and the non-cooperative game method. Then, the second MPMAB problem is to determine the best SF for each IoT device by using the Thompson sampling (TS) algorithm. Thereafter, the optimal RIS and SF can be obtained by alternately iteratively running the two-stage MPMAB problem. Finally, the theoretical result provides an upper performance bound for the E2Boost algorithm. In addition, the numerical results show that the performance of the proposed algorithm is improved by M times compared with the existing distributed allocation strategy, where M is the number of SFs of each IoT device.

(3) To deal with the age-of-information (AoI) based scheduling problem in high-density IoT networks, this thesis proposes a generalized Whittle Index-based scheduling strategy by combining the Markov decision process (MDP) and the MAB technology. First, this AoI-based minimization problem is formulated as a correlated restless MAB (CRMAB) by considering the correlation among IoT devices. Then, this CRMAB problem is decoupled into several one-dimensional subproblems by using the decomposition theory. In the stochastically identical channel model, this thesis derives the closed-form expression of the generalized Whittle index (GWI) by solving the Bellman equation of each sub-problem and proposes a GWI-based scheduling strategy. In the stochastically non-identical channel model, this thesis derives the closed-form expression of the generalized partial WI (GPWI), and proposes a GPWI-based scheduling strategy. Finally, the theoretical result provides two lower performance bounds for the proposed GWI- and GPWI-based scheduling strategies. In addition, the simulation results show that the proposed scheduling strategies can significant outperform the existing AoI-based scheduling algorithms in high-density networks.

(4) To attack the link adaptation problem in underwater acoustic communication, this thesis proposes a joint transmission frequency and rate selection strategy with fast convergence rate and low computational complexity by combining the model features and the MAB technology. We first model it as a single player MAB framework based on the online learning theory. For the stationary channel model, this thesis proposes a

unimodal objective-based TS (UO-TS) algorithm by using the two-dimensional unimodal feature of the objective function. For the non-stationary channel model, this thesis proposes a hybrid change detection (HCD) based TS (HCD-UO-TS) algorithm to jointly track the slowly varying channel and detect the abruptly changing point. For large action space and lack of the unimodal feature, this thesis proposes an iterative boundary-shrinking (IBS) based TS (IBS-TS) algorithm based on the logistic regression-based action classification model. Finally, the theoretical result provides an upper regret bound for the UO-TS algorithm, and shows that its convergence rate is about $\log_2(MN/5)$ faster than the traditional MAB algorithms, where M and N are the number of the transmission frequencies and data rates, respectively.

Key words: Wireless communication network; resource allocation and optimization; multi-armed bandit; optimization theory

目 录

Contents

图索引

表索引

主要缩略语表

5G	5th Generation Mobile Networks，第五代移动通信技术
6G	6th Generation Mobile Networks，第六代移动通信技术
AoI	Age-of-Information，信息新鲜度
BER	Bit Error Rate，误比特率
BS	Base Station，基站
CRMAB	Correlated Restless MAB，状态相关且不停变化的多臂老虎机问题
CST	Carrier Sensing Threshold，载波侦听门限
CSI	Channel State Information，信道状态信息
CSMA	Carrier Sensing Multiple Access，载波侦听多路访问
CTMN	Continue Time-reversible Markov Network，时间可逆马尔科夫网络
CUSUM	Cumulative Sum Control Chart，累积和控制图法
DESim	Discrete Event Simulator，离散事件仿真器
E2Boost	Exploration and Exploitation Boosting，增强的探索与利用
ERWA	Exponential Recency Weighted Average，指数就近加权平均法
ETC	Explore-Then-Commit，先探索后执行
Exp3	Exponential-weight Algorithm for Exploration and Exploitation，基于指数加权的探索与利用算法
Exp4	Exp3 with Expert，带有专家知识的 Exp3 算法
FD	Full Duplex，全双工
GBA	Gradient Bandit Algorithm，随机梯度赌徒算法
GI	Gittins Index，吉汀斯索引
GLR	Generalized Likelihood Ratio，广义似然率
G-ORS	Graphical-based Optimal Rate Sampling，基于图论的最优

	速率采样算法
GPWI	Generalized Paratial Whittle Index，广义部分惠特尔索引
GWI	Generalized Whittle Index，广义惠特尔索引
HCD	Hybrid Change Detection，混合突变检测
HCD-UO-TS	Hybrid Change Detection and Unimodality Objective based Thompson Sampling，基于混合检测和单峰目标的汤普森采样算法
HD	Half Duplex，半双工
IBS	Iterative Boundary Shrinking，迭代边界收缩
IBS-TS	Iterative Boundary Shrinking Thompsom Sampling，基于迭代边界收缩的汤普森采样算法
IoT	Internet-of-Things，物联网
KS	Kolmogorov-Smirnov，科尔莫哥洛夫-斯米尔诺夫
LAI	Logarithm Access Intensity，对数接入强度
LinUCB	Linear UCB，线性 UCB 算法
LoS	Line of Sight，视距
LoRa	Long Range，一种远距离低功耗传输协议
LPC	Linear Prediction Filter Coefficients，线性预测与滤波系数
LSE	Least Square Extimation，最小二乘估计
MAB	Multi-Armed Bandit，多臂老虎机
MAC	Medium Access Control，媒体访问控制
MDP	Markov Decision Process，马尔科夫决策过程
MPMAB	Mulit-Player MAB，多个玩家的多臂老虎机
NLoS	Non Line of Sight，非视距
OFDM	Orthogonal Frequency Division Mutiplexing，正交频分复用
PI	Partial Index，部分索引
PMF	Probability Mass Function，概率质量函数

PSD	Power Spectral Density，功率谱密度
QPSK	Quadrature Phase Shift Keying，正交相移键控
RERWA	Refined ERWA，改进的指数就近加权平均法
RIS	Reconfigurable Intelligent Surface，可重构智能反射面
RMAB	Restless MAB，状态不停变化的多臂老虎机问题
RRM	Radio Resource Management，无限资源管理
SAO	Stochastic and Adversarial Optimal，随机与对抗均最优
SAO-FD-CSMA	SAO Full Duplex Carrier Sensing Multiple Access，随机和对抗都最优的全双工多路侦听访问算法
SF	Spreading Factor，扩频因子
NR	Signal-to-Interference-plus-Noise Ratio，信噪比
SINR	Signal-to-Interference-plus-Noise Ratio，信干噪比
SW	Sliding Window，加窗法
TP	Transmit Power，传输功率
TS	Thompson Sampling，汤普森采样
UCB	Upper Confidence Bound，上置信区间
UE	User，用户
UO	Unimodality Objective，单峰目标
UO-TS	Unimodality Objective based Thompson Sampling，基于单峰目标的汤普森采样算法
WD	Wasserstein Distance，瓦瑟斯坦距离
WI	Whittle Index，惠特尔索引

第 1 章 绪　论

1.1 研究背景与意义

随着第五代移动通信系统（5th Generation Mobile Networks, 5G）在全球范围内的大规模部署，无线网络正向软件化、虚拟化、智能化、用户超致密化演进；与此同时，虚拟与增强现实（Virtual and Augmented Reality，VAR）、电子商务、全覆盖通信、机器对机器（Machine-to-Machine, M2M）通信以及增强的移动宽带等众多新兴应用和垂直领域的出现[1][2]，使得 5G 并不能满足 2030 年及以后的无线通信网络需求[3][4][5]。目前，6G 的研发已在全球兴起。为实现“一念天地，万物随心”的总体愿景，6G 将在性能指标与应用场景上有更高的要求，比如提供全球覆盖，更高的频谱、能量、成本效率，更强的安全性和泛在的智能化；6G 将依赖于新的使能技术，比如新颖的空口与传输技术、数字孪生、网络安全、智能反射面（Reconfigurable Intelligent Surface，RIS）、全双工（Full Duplex，FD）技术、以及云计算等；6G 将不仅仅局限于陆地无线通信网络，还需要整合卫星、无人机和水下传感器等非陆地网络作为有效补充，构建覆盖全球的空天地海一体化通信网络[6][7]。一方面，全频谱资源将被充分挖掘进一步提升数据传输速率和连接设备密度，包括 Sub-6 GHz、毫米波、太赫兹、光频段等；另一方面，由于超异构网络、多种通信场景、大量天线单元、大带宽、新的服务需求的出现将产生海量数据，6G 还将借助人工智能与大数据技术实现泛在智能[8][9]。

由此可见，下一代无线通信网络将变得更加致密化、异构化以及智能化，这将给无线网络中的资源分配与优化带来一系列挑战[10]。例如，在全双工载波侦听多路访问（Full Duplex-Carrier Sensing Multiple Access，FD-CSMA）网络中，FD 链路存在严重的自干扰①和链路间干扰问题，进而降低网络的空间复用率；在分布式异构物联网中，缺少中心节点的用户需要在完全分布式的情况下，完成最优分配策略，最大化网络整体吞吐量；在时延敏感的物联网中，基于时延或吞吐量的资源调度方

① 在全双工通信中，设备可以同时在同一频段上进行发送和接收操作。因此，天线可能接收到自身发送的信号，该回环信号又称之为自干扰信号。

法已经难以满足用户的服务与质量（Quality-of-Service，QoS）需求；在复杂海洋环境中，水声通信网络的信道状态信息（Channel State Information, CSI）通常很难准确获取，从而无法有效保证链路的传输可靠性。面对这些挑战，传统的基于排队论、优化理论、随机几何模型、博弈理论等方法已无法有效解决上述挑战，需要寻找新的技术或理论方法来解决下一代无线通信网络中的资源分配与优化问题。

另一方面，机器学习（Machine Learning，ML）因其不需要准确的数学模型，且具有实时学习和处理复杂场景的特性，已经被广泛应用于无线通信领域[11][12][13]。传统的 ML 方法通常包含以下三类：监督学习[14]、非监督学习[15]和强化学习[16]（Reinforcement Learning，RL）。其中，监督学习属于任务驱动型，重在对样本或数据的利用，即在给定数据集以及标签的情况下，监督学习需要完成给定的任务，比如训练回归、分类和预测模型；与其相反，非监督学习属于数据驱动型，重在探索，即在给定训练数据的情况下，非监督学习需要学习出数据中隐藏的模式，比如聚合；强化学习介于监督学习和非监督学习之间，通常指玩家通过与其周围环境互动来得到相应的奖励，从而调整后续动作来达到其既定的长期目标。在该过程中，如何权衡利用和探索（Exploitation and Exploration，EE）的困境是强化学习面临的一项重要挑战。当前，由于算力的大幅提升和深度神经网络的快速发展，机器学习方法在通信领域中的应用又迎来了一个爆炸式的增长时期[9]。例如，机器学习方法应用于信号检测[17]、信道编码和解码[18]、信道估计和预测[19]、端到端通信[20]、资源分配[21]和能量收集[22]等方面。但是，这些方法通常面临计算复杂度高、没有严格的理论性能保证、以及高维度诅咒等问题。

由此可见，单纯的基于某一类理论的资源分配与优化策略存在各自的局限性，无法有效解决解决下一代无线通信网络中的资源分配与优化问题，需要结合多种理论或方法，从用户实际 QoS 需求出发，设计高效、稳健的资源分配与优化算法。

近几年，多臂老虎机（Multi-Armed Bandit，MAB）技术因计算复杂度低、易于实现、且有严格的理论性能保证被广泛应用于各领域中。MAB 最早由 William R. Thompson 于 1933 年提出并用于临床研究[23]。在第二次世界大战时，德国科学家利用 MAB 来研究当时被认为重要且具有挑战性的战略物资调度问题。直到 1950 年，MAB 才被 Lai 和 Robbins 建模成一个序贯决策问题，并第一次给出了这类问题的理论分析框架[24]。沉默一段时间之后，MAB 在 1980 年开始被广泛应用于工

程和经济领域。近几年，随着互联网、在线推荐系统以及强化学习技术的快速发展，MAB 又重新引起了人们的关注，成为一个新的研究热点。此外，在互联网中，MAB 也被广泛用于广告、电影推荐和搜索引擎设计等应用。例如，谷歌的蒙特卡洛树搜索。在机器学习领域，由于 MAB 的试错机制使其具有很强的学习和解决问题能力（即如何有效权衡 EE 困境），因此，在讨论 RL 时 MAB 也成为理解该机制的一个重要环节；在无线通信领域，已经有很大一部分工作将 MAB 应用于无线网络中的资源分配与优化问题中。例如，认知无线电中的信道接入问题[25]、高密度网络中的小型基站功率分配问题[26]、分布式异构网络中的资源分配问题[27]、水声通信网络中的中继选择问题[28]、无线通信链路速率自适应问题[29]、以及分布式网络中的联邦学习问题[30]。但是，MAB 在通信领域中的应用尚处于初步阶段，且当前大部分的工作只是简单的将 MAB 技术应用到通信领域中，没有考虑如何与其它方法或无线网络中的资源分配问题进行深入、有机的结合。

综上所述，本论文将围绕下一代无线通信网络中几种典型应用场景下的资源分配与优化问题，结合优化理论、通信理论、马尔科夫决策过程、数据驱动技术和 MAB 技术，提出一系列计算复杂度低、严格理论性能保证且易于实现的资源调度和优化策略，为实现下一代无线通信网络的愿景和技术指标提供理论支撑和技术保障。

1.2 课题研究现状

1.2.1 无线通信网络资源分配与优化

自二十世纪初人类第一次实现无线电通信，无线通信技术获得了飞速发展，改变了人类的工作和生活方式，让世界各地人们之间的沟通与交流更加方便；与此同时，人们对通信的需求又促进了无线通信的发展[31]，每隔十年左右的时间无线通信技术边更迭一代[32]。如今，5G 已经在全球开始大规模商业部署，6G 的理论研究也在全球紧锣密鼓地进行，甚至成为各个国家发展的一项战略。可以预见，未来的人类生活将离不开无线通信[6][9]。

资源是指一切可被开发和利用的物质、能量和信息的总称，它广泛存在于自然界和人类社会中，可以被人类开采并多次转化带来额外的价值财富[31]。而无线通

信资源泛指传输涉及到的媒介，比如频率、功率、时间和空间等。随着日益复杂的国际形势和逐年恶化的全球气候问题，无线网络资源也正变得越来越稀缺；另一方面，社会的飞速发展、新型的应用场景、以及海量的设备和数据都对无线通信提出了更高的要求。如何有效合理地分配和优化有限的物理资源？这不仅是无线通信中亟需解决的问题，也是未来人类生存面临的一大挑战[32]。

无线通信中的资源分配与优化又称为无线资源管理（Radio Resource Management，RRM）[31]，即在给定资源约束的前提下，最大化网络用户的 QoS，比如吞吐量，用户之间的公平性、带宽、误码率、时延、中断概率、网络寿命、信息新鲜度（Age-of-Information，AoI）等。RRM 的基本出发点是在不同用户 QoS 需求，以及信道特性因衰落、干扰等引起变化的情况下，动态地调整通信系统的传输参数、自适应地分配网络资源，从而最大限度地提高资源的利用率[33]。RRM 算法在提高系统性能等方面发挥着着重要的作用。一方面，通信的频率、时间和空间等物理资源是有限的，开发新的频谱和维度资源需要更先进的技术；另一方面，随着无线通信新技术的发展，比如多输入多输出（Multiple Input Multiple Output，MIMO）技术[34]、认知无线电技术[35]、正交频分复用（Orthogonal Frequency Division Multiplexing，OFDM）技术[36]、波束成形技术[37]、FD 技术[38]等，促使人们寻找更好的资源分配与优化策略。随着 5G 的大规模商业部署，资源分配的研究热点也正朝着高密度、低时延、异构化和智能化等方面发展。下面基于不同的出发点，对 RRM 作简要的总结。

首先，从研究技术或手段上来看，RRM 方法通常有基于优化理论[89]、启发式算法[40]、随机几何模型[41]、博弈论[42]和机器学习等方法。同时，根据优化问题的目标（比如最大化吞吐量或传输速率之和、最小化网络平均 AoI，最大化频谱效率、最小化端到端时延，最大化能量效率和最小化丢包率等），其建模和求解方法各不相同。换言之，针对不同的资源分配问题，选择合适的资源分配技术或手段，有助于更快地获得系统的最优性能。

通过合理的假设与近似，上述资源分配与优化问题通常可以描述为一个最优化模型，根据目标函数的凸或非凸特性采取相对应的求解方法。如果问题的目标函数和约束条件均是凸的，可以采用已有的凸优化方法求解，比如 KKT（Karush-Kuhn-Tucker）条件或拉格朗日乘子法求解，通过迭代运算（如梯度下降的方法，

牛顿法等）得到目标函数的最优解；如果该约束问题的目标函数或约束条件是非凸的，可以通过求解拉格朗日对偶问题的解来得到原问题的次优或最优解。此外，对于一些特殊的优化问题可以对其进行近似求解。比如混合整数规划问题中的变量数目较大时，可以通过取整变量的方法得到的近似解，或将其凸松弛之后采用凸优化方法求解来获得最优解的上界或下界。然而，上述求解方法需要假设问题模型固定且已知。当研究对象是一个动态系统时，则需要采用博弈论的方法对系统进行建模。具体地，系统中节点之间的行为互相作用，每一个节点自私而理性地最大化其自己的利益（即其行动不会刻意去损害别人的利益），通过为每个节点设计反映其自身需求的效用函数，不同节点之间进行合作或非合作博弈，直到系统达到动态平衡（即均衡）。但是，系统的均衡态不固定，且均衡的结果也不一定是全局最优。博弈理论的目标是在系统达平衡时，该均衡点即系统的最优性能，而由博弈论归纳出的模型也称为最优化模型。综上，两类方法各有其优缺点：优化方法适合静态系统、但计算复杂度较高；博弈论适用于动态系统、但无法保证达到系统的最优解。此外，针对一些 NP（Non-deterministic Polynomial）或者 NP-hard 问题，通常采用蚁群算法[43]、粒子群算法[44]、遗传算法[45]等启发式搜索算法进行求解，获得近似最优解。另外，如果问题模型未知或优化变量较大时，可以采用机器学习的方法进行求解。

其次，根据不同的网络拓扑结构，比如集中式和分布式，RRM 的实现方式也不同。

在集中式网络中，通常网络中存在一个中心节点（如基站），负责统一收集、传递、协调各子节点的信息[46]。例如，在集中式信道分配策略中，各节点将其信道状态信息上传至基站，基站对收集到的信息进行处理并给每一个子节点分配最佳的信道，从而最大化网络的整体性能。其优点是能够考虑不同节点的需求，且能达到全局最优；缺点是需要额外的中心节点来收集各子节点的信息，这将增加系统负担和通信开销，且信道的时变性又要求算法很强的鲁棒性和容错性。集中式资源分配适合传统的蜂窝网和带有中心节点的无线自组织网络。在分布式网络中，没有中心节点来协调、收集各节点的信息，它们需要互相进行信息传递，获取决策需要的关键参数信息，分布式地进行资源分配[47]。此外，在完全分布式的网络中，节点之间无任何信息传递，各节点通常通过非合作博弈的方式来得到系统的均衡态。例如，

在异构网络中，因为各节点采用不同的协议或支持的功能不同，它们之间通常不存在任何的信息交互。因此，分布式资源分配的优点是较少的通信开销，较低的算法复杂度；缺点是系统需要严格同步、且无法保证全局最优。

再次，根据应用场景不同，RRM 方法朝着新的使能技术和异构网络等复杂场景演进。

在使用 FD 技术的 CSMA 网络中，FD 链路不但存在残余自干扰，而且链路间干扰比 HD 链路更加严重，使得 FD 链路需要更大的空间来确保可靠的传输，这将大大降低 FD-CSMA 网络的空间复用率[48]。在这种情况下，FD-CSMA 网络的性能是否优于 HD-CSMA 网络？是否可以通过调整 FD 链路的物理层层数（如发送功率和载波侦听门限）或 MAC 层参数（如 CSMA 协议中的侦听机制）来提高 FD-CSMA 网络的性能？这些都是引入 FD 技术后，网络资源分配需要重新考虑的问题。在采用 RIS 辅助的蜂窝网络中，通过 RIS 重构出适用于当前信号传输的信道，成倍提高网络传输效率。但是，由于 RIS 自身没有信号处理能力，这将给传统的信道估计带来了新的挑战，即接收端需要联合估计发送端到 RIS 和 RIS 到接收端的 CSI，这对网络中的资源分配算法提出更高的要求。此外，在异构网络场景中，网络从传统的蜂窝小区到宏小区、微小区和微微小区变化，网络中的用户受到的内部干扰以及小区间干扰将更为严重，使得基于网络结构、业务和终端的资源分配问题更为复杂，这也是研究的热点之一。

最后，根据资源分配涉及的协议层不同，RRM 可以在物理层、MAC 层、网络层、传输层以及应用层实现，也可以进行跨层资源分配。在分析全网的性能时，单独针对某一层的优化往往无法有效提高系统的性能。联合优化多层参数，实现跨层优化已经是无线网络优化中的一个主流观点[49]。在下一代无线通信网络中，由于各种新的使能技术和新型应用场景的出现，传统的基于最优化理论的跨层优化面临严重挑战。因此，需要针对新型应用场景，进一步设计高效、合理的跨层资源分配与优化策略。

1.2.2 MAB 技术框架与研究现状

MAB 是一类用于解决在线序贯决策问题的框架，通常被定义在随机调度理论范畴[23]。在过去的几十年中，MAB 技术得到了飞速发展，使其成为一个丰富的多

学科研究领域，受到计算机科学、运筹学、经济学和统计学的青睐。二十一世纪初，随着互联网和在线学习技术的快速崛起，MAB 被广泛应用于互联网中的广告、电影推荐系统、蒙特卡洛树搜索引擎，以及 AlphaGo 等益智游戏中[54]。受此启发，近几年 MAB 在无线通信领域的应用也呈指数增长。下面分别针对 MAB 技术的基本定义、术语、算法、复杂度以及研究趋势作简要阐述。

（1）MAB 技术的基本定义与术语

MAB 定义为在给定有限资源集合的情况下，每一个时隙或回合开始时，玩家必须从该集合中选择一个动作（或臂）与其周围环境进行交互；回合结束后，玩家将获得一个来自环境的反馈（或奖励）；其目的是在最短的时间内找出最佳的动作，从而最大化其长期的累积奖励。在该过程中，玩家面临探索与利用的困境[23]。一方面，玩家需要在每个回合尽可能多的探索不同的动作来最大化其长期奖励，防止错过最佳的动作；另一方面，它需要在每个回合尽可能多的利用当前认为最佳的动作，防止产生更多的性能损失。因此，MAB 问题的一个核心是如何设计有效的动作选择策略来权衡学习过程中的 EE 困境，最大化其长期奖励。

为了衡量 MAB 的选择策略，通常定义累积遗憾（Regret）这一概念来量化交互过程中的性能损失。根据文献[55]，遗憾定义为玩家在某一回合内选择次优动作而非最优动作所导致的性能损失，其累积遗憾可以表示为

$$R(T)=\max_{i=1,\ldots,K}\sum_{t=1}^{T}X_{i,t}-\sum_{t=1}^{T}X_{I_t,t} \tag{1-1}$$

其中，K 和 T 分别表示总的动作数目和时隙数目；$X_{i,t}$ 表示玩家在时隙 t 选择动作 i 时获得的奖励；$I_t\in\{1,\ldots,K\}$ 表示玩家在时隙 t 选择的动作，其获得的奖励为 $X_{I_t,t}$。上式中的奖励 $X_{i,t}$ 和选择策略 I_t 是随机变量，因此，玩家的期望遗憾（Expected Regret）定义为

$$\mathbb{E}\left[R(T)\right]=\mathbb{E}\left[\max_{i=1,\ldots,K}\sum_{t=1}^{T}X_{i,t}-\sum_{t=1}^{T}X_{I_t,t}\right] \tag{1-2}$$

同时，伪遗憾（Pseudo Regret）定义为

$$\bar{R}(T)=\max_{i=1,\ldots,K}\mathbb{E}\left[\sum_{t=1}^{T}X_{i,t}-\sum_{t=1}^{T}X_{I_t,t}\right] \tag{1-3}$$

其中，$\mathbb{E}[\cdot]$表示期望操作。从式（1-2）和（1-3）可以看出，$\bar{R}(T)\le\mathbb{E}[R(T)]$。接着，令$\mu_i(T)=\sum_{t=1}^{T}X_{i,t}/T$表示动作$i$在时隙$T$估计的平均奖励，则玩家的最优动作和奖励分别为$i^*=\arg\max_{i=1,\ldots,K}\mu_i(T)$和为$\mu^*=\max_{i=1,\ldots,K}\mu_i(T)$。因此，伪遗憾可以重写为

$$\bar{R}(T)=T\mu^*-\sum_{t=1}^{T}\mathbb{E}\left[\mu_{I_t}(t)\right] \tag{1-4}$$

遗憾界作为 MAB 技术的一大特点，其在分析算法的性能上发挥着重要作用。为了理论分析方便，通常采用伪遗憾来作为 MAB 算法的性能指标[55]。令$\Delta_i=\mu^*-\mu_i$表示动作i与最优动作之间的平均奖励之差，则式（1-4）可以进一步表示为

$$\bar{R}(T)=\sum_{i=1}^{K}\Delta_i\mathbb{E}[D_i] \tag{1-5}$$

其中，D_i表示直到时间T为止动作i被选中的次数。1985 年，Lai 和 Robbins 给出了随机 MAB 问题的选择策略的一个通用遗憾下界，即

$$\liminf_{T\to\infty}\ \frac{\bar{R}(T)}{\log_2 T}\ge\sum_{i=1}^{K}\frac{\Delta_i}{\mathrm{KL}(\mu_i,\mu^*)} \tag{1-6}$$

其中，$\mathrm{KL}(\cdot)$表示 Kullback-Leibler 散度，$\log_2(\cdot)$表示以 2 为底的对数。该遗憾下界表明任何选择策略都不可能获得比该下界更少的遗憾。从式（1-6）还可以看到$\bar{R}(T)\approx\mathcal{O}(\log_2 T)$，这表明算法的累积遗憾随时间非线性增加，即当$T\to\infty$时，每个回合的遗憾趋于 0，说明算法可以收敛到最优动作。因此，考察一个 MAB 算法是否可行，只需要证明其累积遗憾是否随时间T呈非线性增长。

（2）MAB 技术的主要分类及其对应算法

根据玩家动作上的奖励过程不同，MAB 大致可以分为随机 MAB、对抗 MAB、马尔科夫 MAB 和基于上下文的 MAB 四类。这四类 MAB 问题及其对应的算法和遗憾界如表 1.1 所示。

表 1.1 MAB 技术的分类、经典算法和遗憾界

MAB 分类	常见算法	选择策略	遗憾界
随机 MAB	ETC 算法	$I_t = \arg\max_{i=1,\ldots,K} \hat{u}_i(t)$	$T^{2/3}\mathcal{O}\left(\sqrt{(K\ln T)}\right)^{1/3}$
	ε-贪婪算法	$I_t = \arg\max_{i=1,\ldots,K} \hat{u}_i(t)$	$T^{2/3}\mathcal{O}\left(\sqrt{(K\ln T)}\right)^{1/3}$
	UCB 算法	公式（1-7）	$\mathcal{O}\left(\sqrt{KT\ln T}\right)$
	KL-UCB 算法	公式（1-8）	$\mathcal{O}\left(\sqrt{KT\ln T}\right)$
	GBA 算法	公式（1-9）	无
	TS 算法	公式（1-10）	$\mathcal{O}\left(\sqrt{KT\ln T}\right)$
对抗 MAB	Exp3 算法	公式（1-11）	$\mathcal{O}\left(\sqrt{KT}\right)$
马尔科夫 MAB	Gittins 索引策略	文献[62]（公式 1）	无
	Whittle 索引策略	文献[63]（定义 1）	无
上下文 MAB	Exp4 算法	文献[61]（算法 5）	$\mathcal{O}\left(\sqrt{(2TN\ln K)}\right)$
	LinUCB 算法	文献[64]（算法 1）	$\mathcal{O}\left(\sqrt{(2TN\ln K)}\right)$

其中，随机 MAB 是指动作上的奖励服从某种随机分布，但分布的具体参数值未知。解决随机 MAB 的常见算法有 ETC（Explore-Then-Commit）算法[56]、ε-贪婪算法[57]、UCB（Upper Confidence Bound）算法[58]、KL-UCB（Kullback-Leibler UCB）算法[59]、GBA（Gradient Bandit Algorithm）算法[23]和 TS（Thompson Sampling）算法[60]。在 ETC 算法中，玩家先对每个动作探索 $t_0 = (T/K)^{2/3}\mathcal{O}\left(T\ln T\right)^{1/3}$ 次，其中，$\ln(\cdot)$ 表示以自然数为底的对数；然后选择当前估计的平均值最大的动作 $I_t = \arg\max_{i=1,\ldots,K} \hat{\mu}_i(t)$，作为下一个时隙交互的动作。在 ε-贪婪算法中，每个回合玩家以 $\varepsilon = t^{-1/3}\mathcal{O}\left(K\ln 1/3\right)^{1/3}$ 的概率随机探索和 $(1-\varepsilon)$ 的概率选择当前估计的平均值最大的动作。在 UCB 算法中，每个回合玩家根据以下公式选择下一回合交互的动作

$$I_t = \arg\max_{i=1,\ldots,K} \hat{\mu}_i(t) + c\sqrt{\frac{\ln t}{D_{i,t}}} \tag{1-7}$$

其中，c 是一个正的常数；$D_{i,t}$ 表示动作 i 在时间 t 内被选中的次数。表达式（1-7）主要包括两项，第一项负责利用，即估计的平均奖励随着时间增加而逐渐收敛到真实的期望；第二项负责探索，即估计偏差随着时间增加逐渐减小。在 KL-UCB 算法中，每个回合玩家根据以下公式选择交互动作

$$I_t = \arg\max_{i=1,\ldots,K} \left\{ \underbrace{\max_{\mu_i} \left\{ D_{i,t} \mathrm{KL}\left(\hat{\mu}_i(t), \mu_i\right) \le \ln t + c \ln\ln t \right\}}_{\text{KL-UCB 索引}} \right\} \tag{1-8}$$

值得注意的是，上式中的 KL-UCB 索引需要利用牛顿法或二分法得到。在仿真中，KL-UCB 算法的性能通常优于 UCB 算法，这是因为利用 Kullback-Leibler 散度在限定 $\hat{\mu}_i(t)$ 的上界时比 UCB 算法中的二次方更为严苛，即 Pinsker 不等式 $\mathrm{KL}\left(\hat{\mu}_i(t), \mu_i\right) \ge 2\left(\hat{\mu}_i(t) - \mu_i\right)^2$。在 GBA 算法中，玩家首先给每一个动作分配一个参数 H_i，接着利用 Soft-Max 方法计算每个动作可能被选中的概率 P_i，即

$$P_i = \frac{\exp\left(H_i\right)}{\sum_{k=1}^{K} \exp\left(H_k\right)} \tag{1-9}$$

最后根据该概率得到当前被选中的动作 I_t，并更新 H_i。在 TS 算法中，玩家首先从每一个动作的先验分布 $P_{\mu_i}(t)$ 中产生一个随机值 $\hat{\mu}_i(t)$，再利用下式选择动作

$$I_t = \arg\max_{i=1,\ldots,K} \hat{\mu}_i(t) \tag{1-10}$$

最后，根据交互得到的奖励 $X_{I_t,t}$ 更新被选择的动作 I_t 的后验概率，即 $P_{\mu_{I_t}}(t+1) \propto P\left(X_{I_t,t} \mid \mu_{I_t}\right) P_{\mu_{I_t}}(t)$，其中，$P\left(X_{I_t,t} \mid \mu_{I_t}\right)$ 是似然函数。

对抗 MAB 是指动作上的奖励由对抗者给定，玩家的选择策略是与对抗者进行零和博弈[61]。解决对抗 MAB 的经典算法是 Exp3（Exponential-weight Algorithm for Exploration and Exploitation）算法[61]，其主要思想是利用估计的奖励来更新动作上的参数 H_i，然后利用 H_i 计算每个动作可能被选中的概率 P_i，即

$$P_i = (1-\lambda) \frac{\exp\left(H_i\right)}{\sum_{k=1}^{K} \exp\left(H_k\right)} + \frac{\lambda}{K} \tag{1-11}$$

其中，$\lambda \in (0,1)$，最后根据该概率得到当前被选择的动作 I_t，并更新 H_i。从上式可以看到，第一项负责利用，即跟随最有可能获得最大奖励的动作；第二项负责探索，即随机对 K 个动作进行探索。Exp3 算法的主要优点是在最坏情况下仍可以保证其遗憾上界是非线性的，且具有较低的计算复杂度。

马尔科夫 MAB 是指每一个动作有选中和未选中两个状态，动作上的状态和奖励变化服从马尔科夫转移过程。当玩家在每一个回合只选择一个动作 $M=1$，且未被选中的动作的状态不发生变化，则称为 Rest MAB 问题；当玩家每一个回合选择不止一个动作 $M>1$，且未被选中的动作的状态也会发生变化，则称为 Restless MAB 问题。对于 Rest MAB 问题，其经典求解算法是基于 Gittins 索引[62]的调度策略；对于 Restless MAB 问题，其经典求解算法则是基于 Whittle 索引[63]的调度策略。两种索引策略的核心思想是：在每一个回合，玩家选择最大的 $M=1$ 个 Gittins 索引或 $M \geq 1$ 个 Whittle 索引对应的动作与环境进行交互。两种索引策略的主要优点是计算复杂度低，且都渐进最优。

基于上下文的 MAB 是指玩家在决策前可以观测到一个附加信息（比如玩家所处的环境信息和任何有助于决策的信息），并在该信息下做出最佳决策。上下文 MAB 类似于强化学习，同样面临复杂的环境（因为附加信息可以看成不同的系统状态），且有很广的应用，比如 Google 搜索和广告推荐系统。对于上下文 MAB 问题，其经典求解算法主要有 Exp4（Exponential-weight Algorithm for Exploration and Exploitation with Expert）算法[61]和 LinUCB（Linear UCB）算法[64]。前者适用于奖励具有对抗特点的 MAB 问题；而后者适用于奖励是关于上下文特征的线性函数的 MAB 问题。值得注意的是，表 1.1 中的 N 指总的上下文的数目。Exp4 算法的核心思想是在 Exp3 算法基础上将上下文看成专家，然后根据专家的建议和估计奖励来更新参数 H_i；而 LinUCB 算法的核心思想是在计算每一个动作的 UCB 索引时，需要联合考虑上下文特征和奖励。两种算法的优点都是有严格的性能保证，即遗憾的上界。

此外，根据玩家的数目不同，MAB 还可以分为单个玩家的 MAB（Single-Player MAB，SPMAB）问题和多个玩家的 MAB（Multi-Player MAB，MPMAB）问题。SPMAB 问题已在前面阐述，下面简要介绍 MPMAB 问题。通常，MPMAB 又可以

分成集中式和分布式两种。在集中式 MPMAB 问题中，所有玩家共享相同的动作空间或集合，且有一个中心节点来协调各个玩家之间的选择策略。例如，在采用 UCB 算法的 MPMAB 问题中，中心节点可以根据各玩家的 UCB 索引来调整其选择策略，防止不同的玩家选择相同的动作而产生碰撞或挤兑。分布式 MPMAB 问题是当前研究的一个热点。其核心问题是：当多个玩家参与决策时，如何采用 MAB 选择策略在完全分布式的情况下，使得系统的整体性能达到最优？通常分布式 MPMAB 问题有以下三种设定：（1）每一个玩家有不同的动作空间，且多个玩家选择相同的动作不会产生碰撞；（2）所有玩家共享相同的动作空间，但多个玩家选择相同的动作将产生碰撞，得到奖励均为零；（3）所有玩家共享相同的动作空间，多个玩家选择相同的动作不会产生碰撞。针对第一类分布式 MPMAB 问题，文献[65]提出了一种在随机和对抗环境中都最优的 MAB 算法；针对第二类分布式 MPMAB 问题，文献[66]和[67]分别提出了 Musical Chair 算法和 GoT（Game of Throne）算法；最后，针对第三类分布式 MPMAB 问题，文献[68]提出了分布式的 ETC 算法。

（3）MAB 算法的复杂度和实际应用考量

下面简要分析 MAB 算法的复杂度和其在实际应用中的主要考量。

一个算法的复杂度通常包含时间复杂度和计算复杂度两部分。就算法的收敛速度而言，时间复杂度和计算复杂度可以看作同一个概念，即研究当问题规模变大时算法执行时间增加的趋势。在上述 MAB 算法中（不包括 KL-UCB 算法），每个回合算法只执行基本运算操作，算法的计算复杂度为$\mathcal{O}(T)$，属于多项式时间复杂度。但是在一些具有特定结构的 MAB 问题中（如上一节中的随机 MAB 问题），已知的低遗憾算法（如 KL-UCB 算法和 TS 算法）通常具有较高的计算复杂度，因此对于大型的问题，必须依赖能够在合理时间内运行的简单算法（如 ETC 算法和ε-贪婪算发）。因此，MAB 在实际应用中的主要考量是：如何设计一个 MAB 算法来有效权衡其计算复杂度和伪遗憾上界？

（4）MAB 的国内外研究现状

关于 MAB 的研究现状大致可以分为理论研究和应用研究两方面。

在理论研究方面，目前主要是基于传统 MAB 框架构造不同的 MAB 变型问题，再针对各种变型的 MAB 问题设计相应的算法，分析和推导所提算法的遗憾上界。

这种变型包括奖励过程、动作空间特征以及算法的目标等，下面结合不同的变型阐述 MAB 技术在理论方面的研究现状。

传统的随机 MAB 问题通常假设奖励过程是平稳的。文献[69-75]考虑动作上的期望奖励存在缓慢变化或突然变化的情况，研究了奖励过程非平稳的随机 MAB 问题。针对缓慢变化的奖励过程，文献[69]和[70]提出了一种样本滑动窗口（Sliding Window，SW）方法，即在计算动作的平均经验奖励时只考虑时间长为 SW 的最近的奖励；同样考虑缓慢变化的奖励过程，文献[71]和[72]提出了一种指数新近加权平均(Exponential Recency Weighted Average，ERWA)方法来计算经验平均奖励，即赋予最近的奖励较大的权重，过去的奖励较小的权重。突然变化的奖励过程指各动作上的平均奖励在一个时间块内保持不变，但在块之间快速变化，且这种变化无法通过 SW 方法和 ERWA 方法来处理。针对突变的奖励过程，文献[73]和[74]基于 CUSUM (Cumulative Sum Control Chart)和 KS（Kolmogorov-Smirnov）断点检测方法分别提出了 CUSUM-TS 算法和 KS-TS 算法。此外，文献[75]基于广义似然率（Generalized Likelihood Ratio，GLR）法提出了 GLR-UCB 算法。

传统的 MAB 问题通常假设动作之间是相互独立的。当考虑动作之间的相关性时，这类问题又被归纳为基于模型的 MAB 框架。文献[76]和[77]考虑聚类 MAB 问题，考虑将动作空间中的相关动作聚合成一个超级动作，然后在该超级动作上执行传统的 MAB 算法。文献[78-80]考虑了单峰 MAB 问题，假设动作空间中的各动作的平均奖励具有单峰结构或拟凹函数特性，然后利用该特性提出了一种基于单峰特性的 MAB 算法。文献[81]研究了具有图结构的 MAB 问题，采用无向图来表征了各动作之间的相关性，然后利用各顶点之间的联系提出了一种基于图模型的 MAB 算法。综上，通过利用动作之间相关性，上述算法可以显著减少玩家的探索时间，从而以加快算法的收敛速度。

传统的 MAB 问题通常假设动作空间是离散的。文献[50]考虑动作空间连续的 MAB 问题，又称为 Lipschitz MAB，是指动作的奖励的均值满足 Lipschitz 条件。对于动作空间连续的 MAB 问题，一种简单的方法是将动作空间离散化，再使用传统的 MAB 算法求解。但这会产生两类误差（即量化误差和算法本身误差），从而无法保证算法收敛到最佳的动作。为了克服这个问题，文献[51]提出了一种自适应离散化机制，又称为 Zooming 算法，其核心思想是在最优动作所在区域设置更多

的探索点，在明显次优的动作区域设置较少的探索点。文献[51]表明 Zooming 算法不仅可以有效逼近最优解，且具有较低的计算复杂度。

最后，MAB 在应用方面的研究主要集中在计算机科学、运筹学、经济学、统计学、工程学以及互联网中的推荐系统。为了聚焦研究课题，下面仅简要介绍 MAB 在无线通信领域中关于资源分配与优化方面的应用。

1.2.3 基于 MAB 的无线网络资源分配与优化

与传统的机器学习方法相比，MAB 具有计算复杂度低、严格的理论性能保证以及易于实现等优点。近几年，MAB 技术被广泛应用于无线通信网络中的资源分配与优化问题。比如，无线通信中的链路自适应问题、以及 CSMA 网络的空间复用问题、集中式和分布式网络的资源分配问题。下面针对这几个方面进行详细阐述。

首先，MAB 应用于链路自适应问题中。文献[82]和[83]利用 MAB 技术研究基于 IEEE 802.11 标准的无线通信系统中联合 MIMO 传输调制和速率选择的问题，结合目标函数的单峰特性和 KL-UCB 算法，提出了一种基于图模型的最优速率采样（Graphical-based Optimal Rate Sampling，G-ORS）算法。文献[84]和[85]考虑了同样的调制和速率选择问题。但是，文献[84]基于目标的单峰特性和 TS 算法，从贝叶斯理论出发提出了一种改进的 TS（Modified TS，MTS）算法，理论分析表明 MTS 算法与 G-ORS 算法拥有相同的遗憾上界；此外，文献[85]基于拒绝采样和序贯逆变换方法，进一步提出了一种带约束的 TS（Constrained TS，Co-TS）算法。另外，MAB 还广泛应用于在通信链路的功率控制问题[27]和水声通信链路的中继选择问题[28]。

其次，MAB 应用于 CSMA 网络的空间复用问题中。文献[86]考虑基于 IEEE 802.11 标准的无线局域网中用户的传输信道、功率和载波侦听门限选择问题，将其建模成一个 MAB 问题，对比了 ε-贪婪算法、UCB 算法、Exp3 算法和 TS 算法在该网络中的性能，发现 TS 算法在该无线网络中具有最佳性能。文献[87]研究 CSMA 协议下，网络中链路之间存在复杂的干扰现象，导致上述 MAB 问题的奖励既不是随机的也不是对抗的。在这种情况下，对于随机 MAB 问题最优的 TS 算法和对于对抗 MAB 问题最优 Exp3 算法均无法取得理想性能，且该干扰现象在网络中的链路数目很大时将更加严重。因此，文献[88]和[89]首次在 Exp3 算法的基础引入随机

EE 机制，进而提出了一种在随机和对抗中均最优的 Exp3++算法。基于 Exp3++算法，文献[89]进一步提出了一种在随机和对抗环境中渐进最佳的 MAB 算法，并推导了所提算法的对数伪遗憾界和多项式伪遗憾界限。然而，文献[90]指出 Exp3 算法只是 INF（Implicitly Normalized Forecaster）算法在使用负熵条件下的一个特例。因此，基于 Tsallis 熵和 INF 算法，文献[91]又提出了一种在随机和对抗环境下均最优的 Tsallis-INF 算法，数值结果表明 Tsallis-INF 算法的性能显著优于 Exp3++算法。基于该算法，通过调整链路的发送功率、载波侦听门限、对数接入强度参数，文献[48]有效解决了 CSMA 网络中的空间复用问题。

最后，MAB 应用于集中式和分布式网络中的资源分配问题中。文献[92-95]研究了基于马尔科夫 MAB 的物联网设备调度问题，其目标是最小化网络的平均信息新鲜度。文献[92]利用物联网设备的信息新鲜度的单调特性，使用递归迭代方法推导了 Whittle 索引的闭式表达式，然后提出一种基于 Whittle 索引的调度策略来解决该信息新鲜度最小化问题。与此类似，文献[93]提出了一种基于 Whittle 索引的去中心化调度策略。然而，文献[94]和[95]利用流体分析技术给出了该调度策略的渐进最优性质。此外，文献[95]考虑非理想信道情况下的物联网设备调度问题，提出了部分索引（Partial Index, PI）的概念,并基于该部分索引提出了一种总加权索引匹配策略，并表明该算法不仅能有效解决物联网设备调度问题，且拥有计算复杂度低、渐进最优的特点。

MAB 在分布式网络中的资源分配问题通常被建模成 MPMAB 问题，研究各个玩家在分布式情况下的最优选择策略。文献[96]采用完全分布式的 MPMAB 框架来研究 LoRa（Long Range）网络中的扩频因子分配问题，提出了一种基于 Exp3 算法的选择策略来解决这个 MPMAB 问题。然而，Exp3 算法中玩家不仅关心自己奖励最大化，而要尽量破坏对手的奖励；因此，Exp3 算法并不能保证每个玩家的策略最优。文献[97]利用完全分布式的 MPMAB 框架来研究自组织网络中的信道选择问题，通过顺序估计信道的状态信息矩阵，作者基于匈牙利算法提出了一种 PAC-MPMAB（Probably Approximately Correct MPMAB）算法。但是，PAC-MPMAB 算法需要玩家之间进行信息交互，导致额外的信令开销。为解决上述分布式信道选择问题，文献[98]和[99]提出了两种具有最优的完全分布式资源调度策略。其中，文献[98]结合 MAB 算法和拍卖算法在完全分布式的情况下最大化网络中所有用户的

总和率；而文献[99]结合 MAB 算法和非合作博弈理论，提出了一种 GoT 算法来最大化网络中所有用户的传输速率之和。由此可见，MAB 可以和各种技术进行深度融合来有效解决实际中的应用问题。

1.3 研究目标与内容

1.3.1 研究目标

本论文的研究目标是以下一代无线通信网络中几种典型应用场景的资源分配与优化问题为切入点，结合优化理论、博弈理论、马尔科夫决策过程、数据驱动方法和 MAB 技术，提出一系列计算杂度低、严格理论性能保证、且易于实现的资源分配与优化策略。具体包括以下四点：

第一，针对 FD 链路存在复杂的自干扰和链路间干扰问题，通过调整链路的发送功率、载波侦听门限和对数接入强度参数来提高 FD-CSMA 网络的空间复用率。通过将该联合优化问题解耦成 MAC 层的调度问题和物理层的参数选择问题，本文分别提出最优 FD-CSMA 算法和随机、对抗均最优的 MAB 算法；最后，通过交替迭代求解这两个子问题，得到 FD 链路的最佳发送功率、载波侦听门限和对数接入强度，最大化网络的平均吞吐量或空间复用率。

第二，针对分布式异构网络中的资源分配问题，通过为每一台物联网设备分配最佳的智能反射面和扩频因子来提高网络中物联网设备的传输速率之和。为了学习出该网络中智能反射面和传输信道的 CSI，将该联合选择问题建模成一个两阶段的 MPMAB 问题。结合博弈理论和 MAB 技术，本文提出一种 E2Boost 算法来最大化网络中物联网设备的传输速率之和。所提算法可以在完全分布式网络中运行，且能够收敛到最佳值。

第三，针对高密度物联网中设备之间存在相关性的调度问题，采用信息新鲜度作为网络性能指标，利用马尔科夫决策过程和 MAB 技术，分别考虑理想和非理想信道情况，提出一种基于 Whittle 索引的 MAB 调度策略。该调度策略不仅考虑设备自身的信息新鲜度和信道特征，还考虑与其相关的设备信息。因此，通过利用设备之间的相关性，所提策略可以有效降低高密度网络中物联网设备的平均信息新鲜度。

第四，针对水声通信中不存在统一的信道模型的问题，利用在线学习方法，为水声通信链路寻找最佳的传输频率和速率。传统的链路自适应方法需要知道信道模型或信道状态信息，然而该信息在复杂海洋环境中很难获取。因此，本文采用在线学习的方法，忽略信道的具体物理状态信息，利用每次传输成功或失败的反馈作为奖励，学习出频率和速率的组合的传输成功概率。为了进一步提高算法的收敛性，本文结合问题模型和学习算法的特征，提出一种新颖的基于模型在线学习方法。所提方法能够快速收敛到最佳的传输频率和速率，最大化链路的传输效率。

1.3.2 研究内容

针对上述研究目标，本论文结合优化理论、博弈理论、马尔科夫决策过程、数据驱动技术、以及 MAB 技术，围绕 FD-CSMA 网络、分布式异构蜂窝网络、高密度物联网、水声通信网络四个典型应用场景，设计针对下一代无线通信网络的资源分配与优化策略。研究内容包含 FD-CSMA 网络空间复用机制、分布式异构网络资源分配策略、高密度物联网设备调度策略和水声通信网络链路自适应机制。

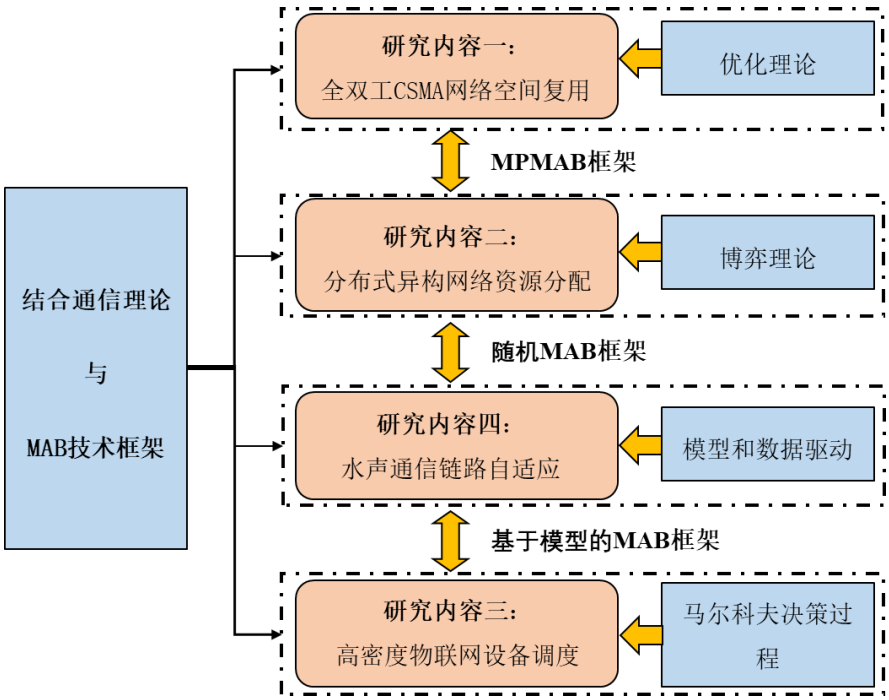


图 1.1 论文主要研究内容与技术路线

本文四个研究内容以及之间的关系如图 1.1 所示。其中，研究内容一和二均涉

及多个玩家之间的交互。因此，可以将其建模成 MPMAB 框架并进行求解。具体地，考虑到研究内容一中玩家之间存在互相对抗的特点，采用对抗 MPMAB 框架对问题进行建模；研究内容二中各玩家自私且理性的最大化自身奖励，因此采用基于博弈论的 MPMAB 框架对问题进行建模。此外，由于研究内容二和四的奖励都服从伯努利分布（即 0-1 奖励）。因此，可以将其建模成随机 MAB 框架，并采用 TS 算法进行求解。不同的是，研究内容二中 TS 算法运行在不同玩家之间，没有利用问题模型的特征；而研究内容四中只有一个玩家，且 TS 算法作为核心算法利用了问题模型的特征。最后，研究内容三和四均采用了基于模型的 MAB 框架，即利用问题模型的特征来提高所提算法的收敛速度。在研究内容三中，本文利用 IoT 设备之间的相关性来设计基于马尔科夫的 MAB 框架，有效地降低网络的平均信息新鲜度；在研究内容四中，本文利用目标函数的单峰特性，设计基于单峰特性的随机 MAB 算法，有效地提高了算法的收敛速度。下面分别对这四个研究内容进行阐述。

（1）基于对抗 MAB 的全双工 CSMA 网络空间复用机制

针对 FD-CSMA 网络中 FD 链路存在复杂的自干扰和链路间干扰的问题，本文研究 FD-CSMA 网络的空间复用问题，联合考虑 FD 链路的功率控制、载波侦听门限调整和对数接入强度自适应，结合优化理论和 MAB 技术，提出了一种在随机和对抗环境下都最优的 FD-CSMA 算法来最大化网络的总吞吐量。首先，将该空间复用问题建模成一个传统的网络优化问题；接着，利用分解理论将其解耦为 MAC 层的调度问题和物理层的参数选择问题。其次，针对这两个子问题分别提出最优 FD-CSMA 算法和随机、对抗均最优的 MAB 算法；然后，通过交替迭代求解上述两个子问题，提出一种在随机和对抗都最优的 FD-CSMA 算法来求解该空间复用问题。最后，理论分析所提算法的收敛性和遗憾上界，并通过数值结果验证了所提算法的有效性。

（2）基于 MPMAB 的分布式异构网络资源分配策略

针对分布式异构蜂窝网络中物联网设备的联合智能反射面和扩频因子选择问题，结合博弈理论和 MAB 技术，本文提出了一种 E2Boost 算法来最大化网络中所有物联网设备的传输速率之和。首先，将该联合选择问题建模成一个传统的组合优

化问题；其次，考虑到物联网设备无法获取智能反射面和信道的准确状态信息，进一步将该问题建模成一个两阶段的 MPMAB 问题，序贯地为每一个物联网学习出设备最优的智能反射面和扩频因子；接着，结合ε-贪婪算法、非合作博弈方法和 TS 算法，提出一种 E2Boost 算法来求解该两阶段 MPMAB 问题；最后，理论分析了所提算法的遗憾上界，且数值结果验证了所提算法的有效性，并表明所提算法的收敛速度不受选择空间大小的影响。

（3）基于马尔科夫 MAB 和信息新鲜度的物联网设备调度策略

针对高密度物联网中基于 AoI 的物联网设备调度问题，考虑不同信道模型和设备之间存在相关性的情况，结合马尔科夫决策过程和 MAB 技术，提出一类基于 Whittle 索引的调度策略来最小化网络的平均 AoI。首先，将该设备调度问题建模成一个马尔科夫决策问题，并通过给每一设备引入一个中间状态，将该高维的马尔科夫决策问题解耦成多个一维的 MAB 子问题；其次，在理想和非理想信道情况下，通过求解每一个子问题的 Bellman 方程，提出了一种基于广义 Whittle 索引的调度策略；接着，通过求解松弛的拉格朗日问题，推导了所提调度策略的性能下界。最后，通过仿真结果验证了所提调度策略的有效性；尤其在高密度网络中，所提调度策略显著优于已有的基于 AoI 的调度策略。

（4）基于随机 MAB 的水声通信链路自适应机制

针对水声通信中不存在统一的信道模型的问题，本文研究水声链路上传输频率和速率的联合选择问题，结合问题模型特征、数据驱动技术和 MAB 技术，提出一类基于模型的 MAB 选择算法来提高水声链路的传输效率。首先，考虑传统的随机 MAB 框架存在收敛速度慢的问题，提出了一种基于模型的 MAB 问题框架；其次，针对平稳信道模型，利用目标函数的二维单峰特性，提出一种基于单峰特性的 MAB 算法；针对考虑非平稳信道的情况，通过联合缓变信道追踪和突变信道检测，提出了一种基于混合非平稳检测的单峰 MAB 算法；针对动作空间较大且目标函数单峰特征不存在的情况，利用传输频率和速率之间的关系来构建动作空间的逻辑回归模型，提出一种迭代边界收缩的 MAB 算法。最后，理论分析所提算法的遗憾上界，并通过数值结果验证了所提算法的有效性。

1.4 论文结构与安排

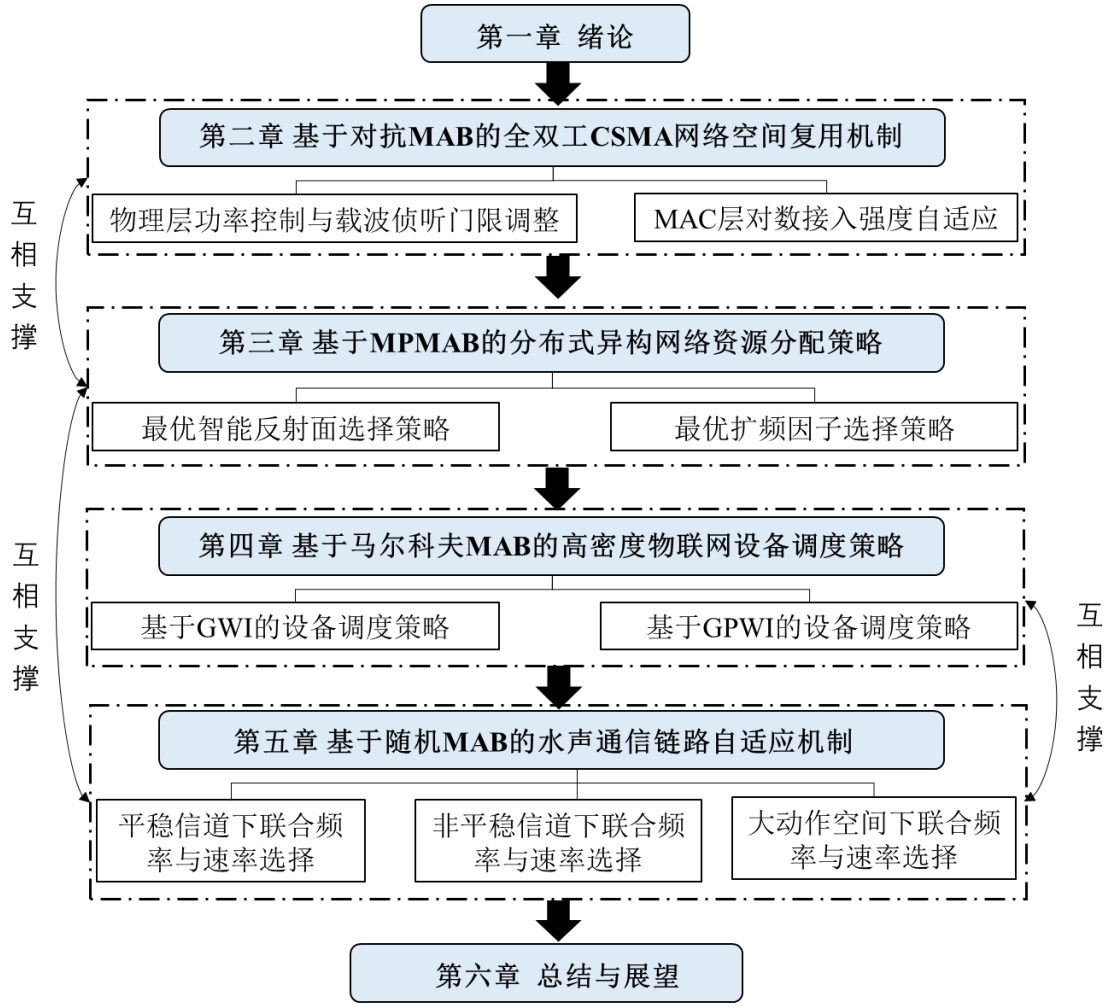


图 1.2 论文研究思路与结构安排

本文共六章，具体结构安排如图 1.2 所示。

第一章为绪论。首先，介绍了论文的研究背景与意义；其次，从无线通信网络中资源分配与优化、MAB 理论框架与算法以及基于 MAB 的资源分配策略三个方面简要回顾了国内外研究趋势；然后，总结了本文的研究目标和内容；最后给出了本文的章节安排。

第二章研究了基于对抗 MAB 的全双工 CSMA 网络空间复用机制。首先，介绍了全双工链路的信号模型和 CSMA 网络的载波侦听模型，并给出了网络在连续马尔科夫模型下的吞吐量计算方法；其次，将该问题建模成一个网络优化问题，并将其分解成 MAC 层的调度问题和物理层的参数选择问题；接着，针对每个子问题，

分别提出了最优的全双工 CSMA 算法和随机对抗均最优的 MAB 算法，并通过交替迭代求解两个子问题得到了主问题的近似最优解；最后，通过理论分析和数值结果验证所提算法的有效性。

第三章研究了基于 MPMAB 的分布式异构网络资源分配策略。首先，介绍了智能反射面辅助的信道模型和扩频因子模型；其次，将该问题建模成一个组合优化问题，并进一步将其建模成两阶段的 MPMAB 问题；接着，提出了一种增强探索和利用的 MAB 算法来求解该两阶段 MPMAB 问题；最后，通过理论分析和数值结果验证了所提算法的有效性。

第四章研究了基于马尔科夫 MAB 和信息新鲜度的物联网设备调度策略。首先，介绍了物联网设备的信道模型、信源相关模型以及信息新鲜度模型；其次，将该问题描述为一个马尔科夫决策问题，并进一步将其分解成多个一维的 MAB 子问题；接着，考虑理想和非理想信道模型，通过求解这些子问题，提出了两种基于 Whittle 索引的 MAB 调度策略；最后，通过理论分析和数值结果验证了所提调度策略的有效性。

第五章研究了基于随机 MAB 的水声通信链路自适应机制。首先，介绍了水声通信的信道模型；其次，提出了一种基于模型的随机 MAB 框架；接着，针对平稳信道、非平稳信道以及动作空间较大的情况，分别提出了三种基于模型的 MAB 算法；最后，通过理论分析和数值结果验证了所提算法的有效性。

第六章对全文的研究内容进行了总结，并对下一步的工作计划进行了展望。

第 2 章 基于对抗 MAB 的全双工 CSMA 网络空间复用机制

本章介绍基于对抗 MAB 的全双工 CSMA 网络空间复用机制[①]。首先，给出全双工信号模型、CSMA 载波侦听模型和基于连续时间马尔科夫模型的网络吞吐量计算表达式。然后，将该空间复用问题分解成 MAC 层的调度问题和物理层的参数选择问题。其次，针对两个子问题分别提出两种的分布式算法，并通过交替迭代求解这两个子问题来最大化网络的空间复用率。最后，理论分析所提算法的遗憾上界，并利用数值仿真结果来验证所提算法的有效性。

2.1 引言

带有碰撞避免机制的载波侦听多路访问（Carrier-Sensing Multiple Access with Collision Avoidance, CSMA/CA）协议能够有效协调同一信道上多个用户的传输，已成为分布式通信系统中广泛采用的 MAC 协议。传统的 CSMA 网络通常建立在半双工（Half Duplex, HD）传输模式上。近几年，全双工（Full Duplex, FD）技术使节点能够在同一频段、同一时间实现接收和发送，被认为能成倍提高通信链路传输速率的一种新技术[8]。

当前，关于 FD-CSMA 网络的性能的研究已经引起人们的广泛关注。与传统的 HD-CSMA 网络相比，FD-CSMA 网络有两个主要特点：第一，严重的残余自干扰。这是导致 FD 链路无法使其传输速率倍增的主要原因。而且，当链路发射功率（Transmit Power, TP）较高时，残余自干扰对链路的影响会更加严重[100]。第二，较大的空间干扰范围。FD-CSMA 网络中 FD 链路的空间干扰范围比 HD-CSMA 网络要大得的多。这是原因并发传输和接收的 FD 链路将产生更大的载波侦听范围和干扰范围，从而导致网络的空间复用率降低（即在同一区域允许同时传输的链路数目减少了）。而且，当网络密度较高时，这种影响将更加严重。

本章研究 FD-CSMA 网络的空间复用问题，通过考虑 FD 链路的 TP 控制、载波侦听阈值（Carrier Sensing Threshold, CST）调整以及对数接入强度（Logarithm

[①] 本章内容已发表于 IEEE Internet of Things Journal 和 IEEE GLOBECOM 2019.

Access Intensity, LAI）自适应来提高网络的总体吞吐量。首先，TP 控制可以减少链路间干扰；其次，CST 调整可以增加并发传输链路的数量；最后，调整 LAI 可以改变链路的 CSMA 参数，从而优化每条链路在单位时间内传输的比例。然而，改变链路的 TP 和 CST 会影响链路的载波侦听关系，使得该网络优化问题具有较高的计算复杂度。另一方面，在一个完全分布式的网络中，该空间复用问题涉及到 MAC 层和物理层，需要进行跨层联合优化这三个参数。为了克服以上挑战，本章首先将该网络优化问题分解为两个子问题，即 MAC 层的联合控制与调度问题，物理层的参数选择问题。其次，针对子问题一，提出一种基于次梯度下降的最优 FD-CSMA 算法，得到 FD 链路最优的 LAI 参数；针对子问题二，提出一种基于对抗和随机均最优（Stochastic and Adversarial Optimal，SAO）的 MAB 算法，得到最佳的 TP 和 CST 参数；最后，通过交替迭代求解这两个子问题，提出 SAO-FD-CSMA 算法来最大化网络的空间复用率。仿真结果验证了所提算法的有效性；而且，与随机调度策略相比，所提算法的性能在静态网络场景和动态网络场景分别提升了约 48%和 43%。

2.2 系统模型

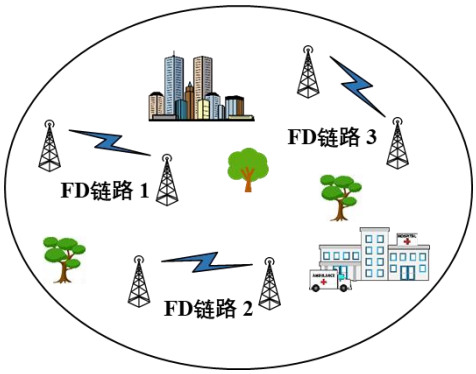


图 2.1 全双工 CSMA 网络的示意图

考虑一个 FD-CSMA 网络，其有 K 个 FD 链路均匀分布在一个区域中，如图 2.1 所示。链路的发射端（Tx）与接收端（Rx）均使用 FD 通信模式，即发送端和接收端可以同时在同一频段上进行发送和接收操作；这些 FD 链路共享同一频带，且采用 CSMA 协议来协调它们之间的传输，即链路采用载波侦听和随机退避机制

进行信道竞争。假设时间离散化为$t=1,2,\ldots,T$。

2.2.1 信号模型

假设不同的 FD 链路采用不同的传输功率，且上、下行信道是对称的。令P_k表示 FD 链路k在 Tx 与 Rx 端的发送功率，其接收信号可以分别表示为

$$\begin{cases} Y_{\mathrm{Rx}_k} = X_{\mathrm{Tx}_k} + SI_{\mathrm{Rx}_k} + AI_{\mathrm{Rx}_k} + n, & \mathrm{Tx}_k \to \mathrm{Rx}_k \\ Y_{\mathrm{Tx}_k} = X_{\mathrm{Rx}_k} + SI_{\mathrm{Tx}_k} + AI_{\mathrm{Tx}_k} + n, & \mathrm{Rx}_k \to \mathrm{Tx}_k \end{cases} \tag{2-1}$$

其中，X_{Rx_k}和X_{Tx_k}分别表示链路k在 Tx 与 Rx 端的传输信号，且服从均值为0、方差为$P_k G_{\mathrm{Tx}_k,\mathrm{Rx}_k}$的独立同分布的高斯分布，其中$G_{\mathrm{Tx}_k,\mathrm{Rx}_k}$表示信道增益。此外，$n$表示标称功率为$\sigma_n^2$的背景噪声。对于 FD 链路，$SI_{\mathrm{Tx}_k}$和$SI_{\mathrm{Rx}_k}$分别表示在 Tx 和 Rx 端接收到的自干扰信号。假设 Tx 和 Rx 端具有相同的自干扰消除能力。根据文献[101]，SI_{Tx_k}和SI_{Rx_k}是服从均值为0、方差为$\chi_k P_k$的瑞利分布，其中χ_k表示链路k的自干扰抑制因子。另外，AI_{Tx_k}和AI_{Rx_k}分别表示在 Tx 端和 Rx 接收到的累积干扰信号，其干扰功率可以分别表示为

$$\sigma_{I,\mathrm{Tx}_k}^2 = \sum_{j\in\Lambda, j\neq k} P_j \left(G_{\mathrm{Tx}_k,\mathrm{Rx}_j} + G_{\mathrm{Tx}_k,\mathrm{Tx}_j} \right) \tag{2-2}$$

和

$$\sigma_{I,\mathrm{Rx}_k}^2 = \sum_{j\in\Lambda, j\neq k} P_j \left(G_{\mathrm{Rx}_k,\mathrm{Rx}_j} + G_{\mathrm{Rx}_k,\mathrm{Tx}_j} \right) \tag{2-3}$$

其中，P_j表示链路j的发送功率；集合Λ表示链路k的干扰范围内的其它链路。最后，链路k在 Tx 端与 Rx 端的接收信干噪比（Signal-to-Interference-plus-Noise Ratios, SINR）分别为

$$\gamma_k^{\mathrm{tx}} = \frac{P_k G_{\mathrm{Tx}_k,\mathrm{Rx}_k}}{\chi_k P_k + \sigma_{I,\mathrm{Tx}_k}^2 + \sigma_n^2} \tag{2-4}$$

和

$$\gamma_k^{\mathrm{rx}} = \frac{P_k G_{\mathrm{Tx}_k,\mathrm{Rx}_k}}{\chi_k P_k + \sigma_{I,\mathrm{Rx}_k}^2 + \sigma_n^2} \tag{2-5}$$

然而，实际通信系统中，根据信号调制编码方式和接收 SINR，每条链路仅支撑有限多个传输速率。令 $\mathbf{C}_k$ 表示链路 k 上所有可能的传输速率集合，并按升序排列为 $c_1 < c_2 < \cdots < c_{|\mathbf{C}_k|}$。因此，每个传输速率将对应于一个 SINR 区间。当接收到的 SINR 落于某一区间时，链路采用对应的速率进行数据传输。值得注意的是，由于链路 Tx 端与 RX 端接收到的干扰信号不同，$\mathrm{Tx}\rightarrow\mathrm{Rx}$ 和 $\mathrm{Tx}\leftarrow\mathrm{Rx}$ 路径上的传输速率也可能不同。假设 $\mathrm{Tx}\rightarrow\mathrm{Rx}$ 和 $\mathrm{Tx}\leftarrow\mathrm{Rx}$ 路径上的传输速率分别为 c_i 和 c_j，则链路 k 的总传输速率为 $\upsilon_k = c_i + c_j$。

2.2.2 载波侦听模型

根据 IEEE 802.11 标准中的 CSMA 协议，链路在开始传输之前需要对信道状态进行侦听。在等待帧间间隔之后，链路随机地从 $[0, CW]$ 内选择一个整数进行退避，其中，CW 是指最初设置为 $CW_{\min}$ 的竞争窗口。每个时隙，若信道的侦听结果为空闲，则退避计数器减1。当退避计数器减少到0 时，该节点方可进行数据传输。若在退避过程中检测到信道处于占用状态，则停止计数，直到再次检测到信道空闲时恢复计数。

然而，FD-CSMA 网络的载波侦听情况与 HD-CSMA 网络不同。首先，考虑两条 FD 链路的情况，即链路 i 和 j。下面，采用成对载波侦听模型来描述它们之间的载波侦听关系。具体地，若两条链路满足以下关系

$$P_j\left(G_{\mathrm{Tx}_i,\mathrm{Rx}_j} + G_{\mathrm{Tx}_i,\mathrm{Tx}_j}\right) \geq S_i \tag{2-6}$$

则认为链路 i 可以侦听到链路 j。在多条链路的网络中，可以定义无向图 $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ 来描述链路间的载波侦听关系。其中，$\mathcal{V}$ 表示 FD 链路构成的顶点，$\mathcal{E}$ 表示两顶点之间的边。图中的边可以表示成一个 $K \times K$ 的矩阵，即 $\mathcal{E} = [\mathbf{e}_1; \mathbf{e}_2; \ldots; \mathbf{e}_K]$，其中，$\mathbf{e}_k = [e_{k,1}, e_{k,2}, \ldots, e_{k,K}]$ 表示矩阵中第 k 行的元素。若 $e_{i,j} = 1$，表示链路 i 可以侦听到链路 j；否则，$e_{i,j} = 0$。此外，假设两条链路上的侦听结果是对称，即 $e_{i,j} = e_{j,i}$。

在该无向图中，网络的传输状态可以由一些顶点组成的集合来定义。在该集合中，任意两个顶点都不存在使其连接的边。假设图 $\mathcal{G}$ 中有 N 个可行的传输状态，记为 $\mathbf{F} = \{\mathbf{f}^i, i = 0, 1, \ldots, N-1\}$，其中，$\mathbf{f}^i = \{f_k^i, k = 1, 2, \ldots, K\}$ 表示第 i 个可行的状态。如

果 $f_k^i=1$，则表示链路 k 可以在第 i 个状态下进行数据传输；否则 $f_k^i=0$。值得注意的是，K 条链路的网络通常有 2^K 个状态，但由于链接之间的竞争关系，可行的网络状态的数量会远小于这个值。

2.2.3 时间可逆马尔科夫模型

为了计算网络的吞吐量，本节将上述载波侦听关系建模成一个连续时间可逆马尔科夫模型（Continued Time-reversible Markov Network，CTMN）。假设网络中的 CSMA 协议是理想的，即退避过程之中或之后链路的传输都无碰撞发生。因此，只需要计算出网络中所有可行状态的平稳分布，便可得到网络的归一化吐量。

令 T_{cd} 和 T_{tr} 分别表示 CSMA 协议的退避值和一个数据包传输的时间，均是服从任意分布的随机数。在 CTMN 模型中，链路的生成率和消亡率分别为 $\lambda=1/\mathbb{E}[T_{\text{cd}}]$ 和 $\mu=1/\mathbb{E}[T_{\text{tr}}]$，其中，$\mathbb{E}[\cdot]$ 表示期望运算。此外，定义链路的对数接入强度（LAI）为平均传输时间与平均退避时间之比，即

$$\rho=\log\left(\frac{\mathbb{E}[T_{\text{tr}}]}{\mathbb{E}[T_{\text{cd}}]}\right)=\log\left(\frac{\lambda}{\mu}\right) \tag{2-7}$$

其中，$\log(\cdot)$ 表示以自然数为底的对数。可以看到，LAI 越大，链路在单位时间内的传输时间越长。

令 λ_k 和 μ_k 分别表示链路 k 的生成率和消亡率。在给定各链路 LAI 的情况下，网络中所有的可行状态的平稳分布为[143]

$$\pi_{\mathbf{f}^i}=\frac{1}{B_r}\exp\left(\sum_{k=1}^{K}f_k^i\rho_k\right),i=0,1,\ldots,N-1 \tag{2-8}$$

其中

$$B_r=\sum_{j=0}^{N-1}\exp\left(\sum_{k=1}^{K}f_k^j\rho_k\right) \tag{2-9}$$

上式中的第一个求和符号作用于整个可行状态集合 $\mathbf{F}$。因此，链路 k 在可行状态 $\mathbf{f}^i$ 内的吞吐量为 $\upsilon_{k,i}\pi_{\mathbf{f}^i}$。通过对所有的可行状态求和，链路 k 在网络中的总吞吐量为

$$\Gamma_k = \sum_{i=0}^{N-1} \upsilon_{k,i} f_k^i \pi_{\mathbf{f}^i} = \frac{1}{B_r} \sum_{i=0}^{N-1} \upsilon_{k,i} \prod_{k=1}^{K} f_k^i e^{f_k^i \rho_k} \tag{2-10}$$

其中，$\upsilon_{k,i} \in \mathbf{C}_k$ 和 $f_k^i \in \mathbf{F}$ 。

2.2.4 最优化问题模型

下面将该空间复用问题建模成一个优化模型。系统目标是通过调整每条 FD 链路的 TP、CST 和 LAI 来最大化网络的空间复用率或吞吐量，即

$$\begin{aligned} \max_{\rho_k, P_k, S_k} \quad & \sum_{k \in \mathcal{V}} \log \Gamma_k \\ \text{s.t.} \quad & \Gamma_k = \sum_{i=0}^{N-1} \upsilon_{k,i} \pi_{\mathbf{f}^i} \\ & \rho_k \in \mathbb{R}^+, P_k \in \mathbf{P}_k, S_k \in \mathbf{S}_k \end{aligned} \tag{2-11}$$

其中，$\mathbf{P}_k$ 和 $\mathbf{S}_k$ 分别表示链路 k 的发送功率和载波侦听门限集合。目标函数采用 $\log(\cdot)$，是考虑链路之间的比例公平性。

由于变量 P_k 、S_k 、υ_k 和 ρ_k 之间存在隐式关系，这里很难求解得到问题（2-11）的解析表达式。另外，由于 P_k 和 S_k 都是离散取值，导致问题（2-11）的目标函数和可行域都是非凸的。而且，在完全分布式的 CSMA 网络中，没有中心节点来协调链路之间信息。为了克服这些挑战，本章利用分解理论，首先将上述问题分解为两个子问题。其中，子问题一可以看作是传输层拥塞控制和 MAC 层调度的联合问题，即在 TP 和 CST 固定的情况下，优化参数 LAI；而子问题二是物理层的参数选择问题，即在 LAI 固定的情况下，寻找最佳的 TP 和 CST 参数。

2.3 全双工 CSMA 网络跨层优化与调度策略

2.3.1 基于最优 FD-CSMA 算法的 MAC 层优化

首先，通过介绍引理 2.1 来得到子问题一。

引理 2.1. *在固定 TP 和 CST 的情况下，由主问题（2-11）可以分解得到以下问题*

$$\begin{aligned}\max_{\rho_k} \quad & V\sum_{k\in\mathcal{V}}\log\Gamma_k-\sum_{i=0}^{N-1}\pi_{\mathbf{f}^i}\log\pi_{\mathbf{f}^i}\\ \text{s.t.} \quad & \Gamma_k\le\sum_{i=0}^{N-1}\upsilon_{k,i}\pi_{\mathbf{f}^i},\forall k\in\mathcal{V}\\ & \sum_{i=0}^{N-1}\pi_{\mathbf{f}^i}=1,\rho_k\in\mathbb{R}^+,\end{aligned}\tag{2-12}$$

其中，V 是一个正的权重因子。

证明：根据文献[102]，主问题（2-11）等价于以下优化问题

$$\begin{aligned}\max_{\rho_k,P_k,S_k} \quad & \sum_{k\in\mathcal{V}}\log\ \Gamma_k\\ \text{s.t.} \quad & \Gamma_k\le\sum_{i=0}^{N-1}\upsilon_{k,i}\pi_{\mathbf{f}^i}\\ & \sum_{i=0}^{N-1}\pi_{\mathbf{f}^i}=1\\ & \rho_k\in\mathbb{R}^+,P_k\in\mathbf{P}_k,S_k\in\mathbf{S}_k\end{aligned}\tag{2-13}$$

其中，第二个约束条件表示该问题的可行域。因此，当给定各链路的传输功率 P_k 和载波侦听门限 S_k 的情况下，上式可以转化成

$$\begin{aligned}\max_{\rho_k} \quad & V\sum_{k\in\mathcal{V}}\log\Gamma_k-\sum_{i=0}^{N-1}\pi_{\mathbf{f}^i}\log\pi_{\mathbf{f}^i}\\ \text{s.t.} \quad & \Gamma_k\le\sum_{i=0}^{N-1}\upsilon_{k,i}\pi_{\mathbf{f}^i},\forall k\in\mathcal{V}\\ & \sum_{i=0}^{N-1}\pi_{\mathbf{f}^i}=1,\rho_k\in\mathbb{R}^+\end{aligned}\tag{2-14}$$

其中，$\sum_i\pi_{\mathbf{f}^i}\log\pi_{\mathbf{f}^i}$ 表示信息熵，即当所有可行状态的概率相等时，其信息熵最大。通过对比上述两式可以发现，子问题一通过引入一个偏差 $\log(N)/V$，来近似求解问题 $\max\sum_{k\in\mathcal{V}}\log\Gamma_k$。因此，当权重 V 足够大时，这个偏差可以忽略不计。也就是说，在固定 TP 和 CST 的情况下，求解子问题一等价于求解主问题（2-11）。 □

根据文献[103]，容易证明子问题一是一个凸问题。因此，其拉格朗日函数为

$$\begin{aligned}L(\boldsymbol{\Gamma},\boldsymbol{\pi};\boldsymbol{\beta},\eta)=&V\sum_{k\in\mathcal{V}}\log\Gamma_k-\sum_{i=0}^{N-1}\pi_{\mathbf{f}^i}\log\pi_{\mathbf{f}^i}\\&+\sum_{k\in\mathcal{V}}\beta_k\left(\sum_{i=0}^{N-1}\upsilon_{k,i}\pi_{\mathbf{f}^i}-\Gamma_k\right)-\eta\left(\sum_{i=0}^{N-1}\pi_{\mathbf{f}^i}-1\right)\end{aligned}\tag{2-15}$$

其中，β_k 和 η 是对偶变量。然后，KKT 条件可以表示为

$$\sum_{i=0}^{N-1}\upsilon_{k,i}\pi_{\mathbf{f}^i}-\Gamma_k\geq 0,\forall k\in\mathcal{V} \tag{2-16}$$

$$\sum_{i=0}^{N-1}\pi_{\mathbf{f}^i}=1 \tag{2-17}$$

$$\beta_k\geq 0,\forall k\in\mathcal{V} \tag{2-18}$$

$$\beta_k\left(\sum_{i=0}^{N-1}\upsilon_{k,i}\pi_{\mathbf{f}^i}-\Gamma_k\right)=0,\forall k\in\mathcal{V} \tag{2-19}$$

$$\frac{V}{\Gamma_k}-\beta_k=0,\forall k\in\mathcal{V} \tag{2-20}$$

$$-1-\log\pi_{\mathbf{f}^i}+\sum_{k\in\mathcal{V}}\beta_k\upsilon_{k,i}-\eta=0,i=0,1,\cdots,N-1 \tag{2-21}$$

通过求解上述 KKT 条件，得到

$$\eta^*=\log\left(\sum_{i=0}^{N-1}\exp\left(\sum_{k\in\mathcal{V}}\beta_k\upsilon_{k,i}\right)\right)-1 \tag{2-22}$$

和

$$\pi_{\mathbf{f}^i}^*=\frac{\exp\left(\sum_{k\in\mathcal{V}}\beta_k\upsilon_{k,i}\right)}{\sum_{i=0}^{N-1}\exp\left(\sum_{k\in\mathcal{V}}\beta_k\upsilon_{k,i}\right)},i=0,1,\cdots,N-1 \tag{2-23}$$

此外，利用式（2-19）和（2-20），可以得到对偶变量 β_k 的次梯度表达式为

$$\dot{\beta}_k=\left[\frac{V}{\beta_k}-\sum_{i=0}^{N-1}\upsilon_{k,i}\pi_{\mathbf{f}^i}\right]^{\mathrm{Q_r}} \tag{2-24}$$

其中，$\mathrm{Q_r}$ 表示变量 β_k 限定在区间 $[\beta_{\min},\beta_{\max}]$ 内的投影操作。实际上，对偶变量 β_k 和主变量 $\pi_{\mathbf{f}^i}$ 可以分别看作是链路 k 的虚拟队列和可行状态 $\mathbf{f}^i$ 的平稳概率分布。

根据以上分析，下面给出一种分布式最优的 FD-CSMA 算法来求解子问题一，如算法 2.1 所示。其核心思想是：通过记录 FD 链路传输成功的次数（即 $\sum_{i=0}^{N-1}\pi_{\mathbf{f}^i}$）来更新对偶变量 β_k 的次梯度、以及 CSMA 协议的 LAI 参数，从而最大化链路的平均吞吐量。

算法 2.1 最优 FD-CSMA 算法（在链路k上运行）

步骤 1： 初始化参数：CSMA 参数λ_k, μ_k, CW和超参数V, ν, Q_r

步骤 2: 执行$b=1,2,...,B$次以下步骤

步骤 3: 产生均值为$1/\lambda_k$的数据包和均值为$1/\mu_k$的指数分布传输时间

步骤 4: 运行 FD-CSMA 协议

步骤 5: 记录链路传输成功的次数N_r

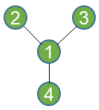
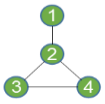
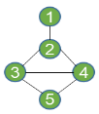
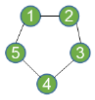
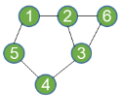
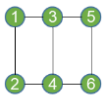
步骤 6： 更新对偶变量β_k： $\beta_k(b+1) \leftarrow \left[\beta_k(b)+\nu(b)\left(V/\beta_k(b)-\upsilon_k N_r\right)\right]^{Q_r}$

步骤 7： 更新 CSMA 参数λ_k和μ_k： $\beta_k=(1/\upsilon_k)\log(\lambda_k/\mu_k)$

从算法2.1可以看到，最优FD-CSMA算法分布式地运行在不同的FD链路上，且按时间段$b=1,2,...,B$进行迭代。在每个时段开始时，链路产生长度为$1/\lambda_k$的数据包和服从均值为$1/\mu_k$的指数分布的传输时间长度。然后，网络执行 CSMA 协议。在该时间段内，链路需要记录下其成功传输的次数N_r。接着，基于该信息，利用梯度下降方法更新对偶变量β_k。最后，根据可行状态$\mathbf{f}^i$和主变量$\pi_{\mathbf{f}^i}$的表达式，可以得到$\beta_k\upsilon_k=\rho_k$，即$\beta_k=(1/\upsilon_k)\log(\lambda_k/\mu_k)$，来更新 CSMA 参数$\lambda_k$和$\mu_k$。经过多次迭代后，对偶变量$\beta_k$将收敛到最佳值，达到求解问题（2-12）的目的。

为了验证所提算法的有效性，本节设计了一种基于 Matlab 平台的 CSMA 离散事件仿真器（Discrete Event Simulator, DESim）。其中，CSMA 协议的参数设置依据 IEEE 802.11g 标准[103]给出，所有结果均由 1000 次蒙特卡洛仿真得到。表 2-1 给出的是在 HD-CSMA 网络中不同网络竞争图下采用 DESim，BoE（back-of-the-envelope）和 CTMN 方法得到的链路归一化吞吐量。其中，竞争图中的每个顶点代表一条 HD 链路，两个顶点之间的连线代表两条链路能互相侦听到对方。此外，BoE 方法为基准方法，由文献[104]给出；而 CTMN 的结果利用公式（2-10）得到。从表 2.1 可以看到，DESin 方法的归一化吞吐量在所有情况下都接近 BoE 方法，且变化趋势与 CTMN 方法一致。因此，仿真结果表明所提 DESin 方法可以有效模拟或实现 CSMA 协议的特征。

表 2.1 在不同网络拓扑下三种仿真方法的 HD-CSMA 网络归一化吞吐量

网络竞争图	 (1)	 (2)	 (3)	 (4)
BoE 方法	（1,0,0,1）	（0,1,1,1）	（1,0,0.5,0.5）	（0.75,0.25，0.25,0.5）
DESim 方法	（0.96,0.01,0.01,0.96）	（0,0.98,0.98,0.98）	（0.98,0,0.60，0.56）	（0.81,0.26,0.31，0.55）
CTMN 方法	（0.74,0.49,0.49，0.75）	（0.18,0.86,0.85 0.85）	（0.95,0.47,0.71 0.72）	（0.67,0.50,0.49 0.66）
网络竞争图	 (5)	 (6)	 (7)	 (8)
BoE 方法	（0.4,0.4,0.4,0.4,0.4）	（1,0,0,1,0,1）	（0.5,0.5,0.5,0.5 0.5,0.5）	（0.2,0.4,0.4,0.8 0.6,0.6）
DESim 方法	（0.47,0.46,0.47 0.46,0.45）	（0.94,0.01,0.02 0.94,0.02,0.94）	（0.50,0.48,0.47 0.51,0.51,0.48）	（0.22,0.49,0.48 0.84,0.66,0.67）
CTMN 方法	（0.61,0.61,0.61 0.60,0.61）	（0.85,0.11,0.11 0.86,0.14,0.86）	（0.85,0.84,0.62 0.62,0.84,0.87）	（0.48,0.60,0.60 0.84,0.75,0.74）

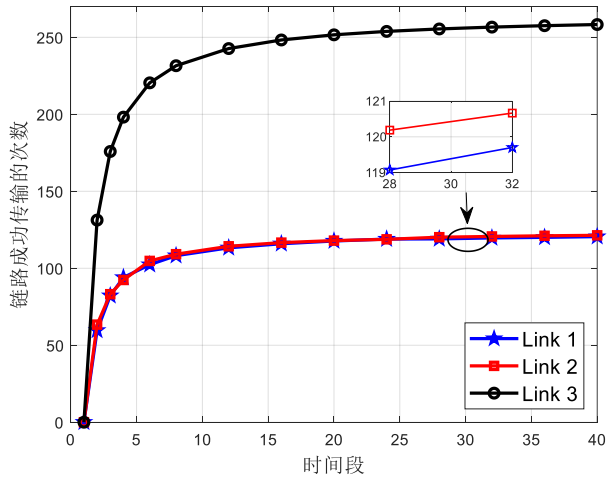


图 2.2 三条 FD 链路的成功传输次数随着时间段变化的曲线

基于上述 DESin 方法，下面考虑一个 3 条 FD 链路的 FD-CSMA 网络场景来验证算法 2.1 的有效性。在该网络场景中，假设链路 1 和 2 能互相侦听到对方。图

2.2 给出的是 3 条 FD 链路的成功传输次数随着时间段变化的曲线，其中，每个时间度的长度 $40\,\mathrm{ms}$；此外，所提算法的参数设置为 $\nu=0.01$，$V=10^6$，$CW=32$，$B_s=40$，$\beta_{\min}=\max\{0,\log(\mu_k^0\log(10/9)/CW)\}$，以及 $\beta_{\max}=\log(\mu_k^0\log 10)$。从图中可以看到，3 条链路的成功传输次数随着时间段的增加而增加，这表明最优 FD-CSMA 算法可以通过自适应调整每条链路的 LAI 参数来最大化网络的吞吐量。此外，由于链路 1 和 2 之间存在竞争关系，它们的性能只有链路 3 的一半左右。

2.3.2 基于对抗 MAB 的物理层调度

在给定 LAI 的情况下，主问题（2-11）可以看成是一个关于物理层参数 TP 与 CST 的联合选择问题，即

$$\begin{aligned}\arg\max_{P_k,S_k}\quad & \sum_{k\in\mathcal{V}}\log\,\Gamma_k\\ \text{s.t.}\quad & \Gamma_k=\sum_{i=0}^{N-1}\upsilon_{k,i}\pi_{\mathbf{f}^i},\forall k\in\mathcal{V}\\ & P_k\in\mathbf{P}_k,S_k\in\mathbf{S}_k\end{aligned}\tag{2-25}$$

根据前面的分析，这是一个非凸问题。而且，调整每条链路的 TP 和 CST 会改变网络中的载波侦听关系和竞争图，导致较高的计算复杂度。因此，考虑采用在线学习的方法来为每条链路学习出最优的 TP 和 CST 参数。在此过程中，需要权衡探索和利用的困境，即每条链路既要充分地探索所有的 TP 和 CST 组合以防错失最佳的 TP 与 CST，又要尽可能多地利用当前最好的 TP 和 CST 对进行传输来最大化吞吐量。因此，子问题二可以建模成一个 MPMAB 框架，其中，玩家是 FD 链路，动作是 TP 和 CST 的组合；奖励是链路的标准化吞吐量，它通过求解子问题一得到。因此，该奖励值在区间 [0, 1] 内变化。每个玩家的目标是自私地最大化其直到时间 T 的累积吞吐量（因玩家之间具有对抗性质，又称对抗 MPMAB 框架）。

下面给出关于 MPMAB 框架的定义与术语。首先，令 $\mathbf{K}=\{1,\ldots,K\}$ 和 $\mathcal{A}_k=\{a_{k,1},a_{k,2},\ldots,a_{k,|\mathcal{A}_k|}\}$ 分别表示所有玩家的集合和链路 k 的所有动作集合。需要注意的是，每个动作是参数 TP 与 CST 的组合。假设每个时隙中包含若干时间段，且在每个时隙内网络的马尔科夫模型都能收敛，则链路 k 选择动作 i 的奖励为

$$r_{a_{k,i}} \doteq \frac{\Gamma_{a_{k,i}}}{\Gamma_k^*} \tag{2-26}$$

其中，$\Gamma_k^* = \max_i \Gamma_k(a_{k,i}), i = 1,\ldots,|\mathcal{A}_k|$ 表示链路 k 在所有动作上可能观测到的最大奖励。链路 k 在时间 T 内的总奖励为

$$R_k \doteq \sum_{t=1}^{T} r_{k,\xi_t^k}(t) \tag{2-27}$$

其中，ξ_t^k 表示当前时隙被选中的动作索引。最后，为了理论分析方便，根据[61]，采用弱遗憾来表征链路 k 的性能，其定义为当前选择的动作与最优动作之间的性能差，即

$$\mathcal{R}eg_k \doteq \max_i \sum_{t=1}^{T} \left(r_{k,i}(t) - r_{k,\xi_t^k}(t) \right), \forall k \in \mathbf{K} \tag{2-28}$$

在该 MPMAB 问题中，总的弱遗憾可以表示为 $\mathcal{R}eg \doteq \sum_{k=1}^{K} \mathcal{R}eg_k$。

算法 2.2 基于 SAO 算法的选择策略（在链路 k 上运行）

步骤 1：	初始化参数：奖励 $\hat{R}_{k,i}(1) = 0, \forall i \in \mathcal{A}_k$ 和经验权重 $\hat{\omega}_{k,i}(1), \forall i \in \mathcal{A}_k$
步骤 2:	执行 $z = 0,1,2,...,$ 次以下步骤
步骤 3:	根据（2-30）和（2-31）分别更新参数 χ_z 和 θ_z
步骤 4:	若 $\max_i \hat{R}_{k,i}(t) \le \theta_z - |\mathcal{A}_k|/\chi_z$，循环执行步骤 5-12；否则，至步骤 3
步骤 5:	利用（2-32）计算所有动作的 PMF $q_{k,i}(t)$
步骤 6：	根据 $q_k(t)$ 选择当前回合的动作 ξ_t^k
步骤 7：	运行基于 DESim 的 FD-CSMA 协议，得到奖励 $r_{k,\xi_t^k}(t)$
步骤 8：	计算每个动作上估计的平均奖励：$\hat{r}_{k,i}(t) = r_{k,i}(t)1_{\xi_t^k = i} / q_{k,i}(t)$, $\forall i \in \mathcal{A}_k$
步骤 9：	计算每个动作上估计的总奖励：$\hat{R}_{k,i}(t+1) = \hat{R}_{k,i}(t) + \hat{r}_{k,i}(t), \forall i \in \mathcal{A}_k$
步骤 10：	利用（2-29）更新每个动作上的权重 $\omega_{k,i}(t)$

由于每条链路只自私地最大化其吞吐量，并不会刻意去破坏其它链路的性能，因此下面提出一种在对抗和随机环境中都最优的 MAB 算法来求解子问题二，简称

SAO 算法，如算法 2.2 所示。从中可以看到，SAO 算法是分阶段进行的，即 $z=0,1,2,...,$。每个阶段的时间长度由估计的总奖励 $\hat{R}_k$ 和指数增长的界限 θ_z 来控制。在每个阶段 z，玩家根据其动作的概率分布 q_k 来选择一个动作（即 TP 和 CST 的组合）。在运行 FD-CSMA 协议后，它将获得一个奖励 $r_{k,\xi_t^k}(t)$，即归一化的吞吐量。然后，根据该奖励计算所有动作的估计的平均奖励 $\hat{r}_{k,i}(t)$ 和总奖励 $\hat{R}_{k,i}$。最后，各个动作上的权重可以更新为

$$\omega_{k,i}(t)=\hat{\omega}_{k,i}(1)\exp\left(\frac{\chi_z}{|\mathcal{A}_k|}\hat{R}_{k,i}(t)\right) \tag{2-29}$$

其中

$$\chi_z=\min\left\{1,\sqrt{\frac{|\mathcal{A}_k|\log(|\mathcal{A}_k|)}{(e-1)\theta_z}}\right\} \tag{2-30}$$

和

$$\theta_z=\frac{|\mathcal{A}_k|\log(|\mathcal{A}_k|)}{(e-1)}4^z \tag{2-31}$$

此外，各个动作上的概率质量函数（Probability Mass Function，PMF）$q_{k,i}(t)$ 可以更新为

$$q_{k,i}(t)=(1-\chi_z)\frac{\omega_{k,i}(t)}{\sum_{j\in\mathcal{A}}\omega_{k,j}(t)}+\frac{\chi_z}{|\mathcal{A}_k|} \tag{2-32}$$

通过上式可以发现：第一，动作的 PMF 依赖于初始权重的 $\hat{\omega}_{k,i}(1)$，因此可以加入网络的先验信息来提高网络的收敛速度；第二，将式(2-31)代入(2-30)，得到 $\chi_z=2^{-z}$，这表明当 z 足够大时，$\chi_z\to 0$，动作的概率分布 $q_{k,i}$ 的第二项趋于 0，SAO 算法只有利用没有探索，即算法处于收敛状态；第三，令 $\hat{R}_{k,i}(t+1)=\sum_{s=1}^{t}\hat{r}_{k,i}(s)$ 和 $\hat{r}_{k,i}(t)=1_{\xi_t^k=i}r_{k,i}(t)/q_{k,i}(t)$，则算法估计的奖励的期望为 $\mathbb{E}[\hat{r}_{k,i}(t)]=r_{k,i}(t)$，以及方差为

$$\mathrm{Var}[\hat{R}_{k,i}(t+1)]=\sum_{s=1}^{t}\left(\frac{1-q_{k,i}(s)}{q_{k,i}(s)}r_{k,i}^2(s)\right) \tag{2-33}$$

其中，Var[·]表示取方差运算。上式表明可以通过增加选择概率来降低估计的奖励的方差，从而得到较小的遗憾上界。

2.3.3 基于 FD-SAO-CSMA 的联合优化与调度

前面已经将主问题分解成两个子问题，分别提出了最优 FD-CSMA 算法和 SAO 算法来求解子问题一和子问题二。下面，通过交替迭代算法 2.1 和算法 2.2 来求解主问题（2-11），简称 SAO-FD-CSMA 算法，如算法 2.3 所示。

从算法 2.3 可以看到，每条 FD 链路异步运行 SAO-FD-CSMA 算法。首先，每条 FD 链路根据动作的概率分布获取一对 TP 和 CST 参数。然后，根据调制方案和接收到的 SINR，物理层确定一个合适的传输速率。接着，MAC 层和传输层共同决定如何激活那些传输速率高、竞争少的链路，以最大化网络吞吐量。具体地，链路通过观察成功传输的次数运行算法 2.1 来得到最优的 LAI。最后，运行算法 2.2 为每个链路寻找到最佳的 TP 和 CST 参数。算法重复上述步骤，直到时间T停止。

算法 2.3 SAO-FD-CSMA 算法（在链路k上运行）

步骤 1： 初始化算法 2.1 和 2.2 中的相关参数

步骤 2: 执行$t=1,2,...,T$次以下步骤

步骤 3: 根据q_k选择一对 TP 和 CST 组合

步骤 4: 根据 SINR 确定链路传输速率υ_k

步骤 5: 运行算法 2.1，得到奖励Γ_k

步骤 6: 利用（2-26）计算链路的归一化吞吐量r_k

步骤 7： 运行算法 2.2，得到权重ω_k和参数χ,θ

步骤 8： 利用（2-32）更新动作的概率分布q_k

2.4 理论分析

本节结合优化理论和 MAB 的有限时间分析技术，分别给出最优 FD-CSMA 算法的收敛性分析和 SAO-FD-CSMA 算法的遗憾上界。由于 SAO-FD-CSMA 算法可以看作最优 FD-CSMA 算法和 SAO 的结合，因此，下面先给出前者的收敛性分析，

再结合 SAO 算法推导后者的遗憾上界。

根据文献[105]和[106]，最优 FD-CSMA 算法可以看作是一个带马尔科夫噪声的随机近似算法。其收敛性分析的难点在于虚拟队列和 CSMA 参数的更新，因为两者都依赖于 FD 链路的随机服务或到达过程。首先，给出如下假设

假设 2.1. 若 $\beta_k^0 \in \mathbb{R}_+, \forall k \in \mathcal{V}$ 和 $\beta_k^0 \upsilon_{k,i} = V \exp\left(\sum_{i=0}^{N-1} f_k^i \pi_{\mathbf{f}^i}\right)$ 已通过求解获得，则有 $\beta^{\min} < \beta_k^0 < \beta^{\max}, \forall k \in \mathcal{V}$ 。

基于假设 2.1，最优 FD-CSMA 算法的收敛性可由定理 2.1 给出，即

定理 2.1. 假设 $\sum_{b=0}^{\infty} \nu(b) = \infty$ 和 $\sum_{b=0}^{\infty} \nu(b)^2 \le \infty$ 。对于任意的 $\beta_k, \forall k \in \mathcal{V}$, 假设 2-1 成立，则最优 FD-CSMA 算法收敛性结果为

$$\lim_{b\to\infty} \beta_k(b) = \beta_k^* \quad 和 \quad \lim_{b\to\infty} \Gamma_k(b) = \Gamma_k^* \tag{2-34}$$

其中，β_k^* 和 Γ_k^* 都是子问题一的解。

证明: 详见文献[103]中定理 1 的证明。

注 1. 定理 2.1 表明通过选择递减的迭代步长，最优 FD-CSMA 算法最终将收敛。换言之，每条链路的平均吞吐量都有一个上限。所以，SAO-FD-CSMA 算法的奖励是有界的，可以归一化到[0, 1]范围内。这一观察有助于推导出 SAO-FD-CSMA 算法的遗憾上界。

注 2. 定理 2.1 表明通过选择合适的超参数 V 可以控制最优 FD-CSMA 算法的收敛速度，即较大的 V 可以提高算法的性能，但会降低其收敛速度。因此，在最优 FD-CSMA 算法的设计过程中，需要仔细权衡性能和收敛速度来选择合适的超参数 V 。

基于定理 2.1，下面推导 SAO-FD-CSMA 算法的遗憾上界。令 $M_k = |\mathcal{A}_k|$ 和 $R_{k,\max} = \max_{j\in\mathcal{A}_k} \sum_{t=1}^{T} r_{k,j}(t)$，则定理 2.2 为

定理 2.2. 对于任意的 $M_k > 0$，$\forall k \in \mathbf{K}$ 和足够长的时间 $T > 0$，有

$$\mathcal{R}eg \le \sum_{k\in\mathbf{K}} \left(8\sqrt{e-1}\sqrt{R_{k,\max} M_k \ln M_k} + 8(e-1)M_k + 2M_k \ln M_k\right) \tag{2-35}$$

该遗憾上界对于任意的分布式 FD-CSMA 网络均成立。

证明： 令 Z_k 表示玩家 k 的总阶段数，$B_{k,z}$ 和 $T_{k,z}$ 分别表示第一个和最后一个阶段内的时隙数目。为了方便分析，将文献[61]中引理 2.2 和 2.3 给出如下：

引理 2.2. 对于玩家 k 上的任意动作和阶段，以下表达式成立，

$$\sum_{t=B_{k,z}}^{T_{k,z}} r_{k,\xi_t^k=i}(t) \geq \sum_{t=B_{k,z}}^{T_{k,z}} \hat{r}_{k,j}(t) - 2\sqrt{e-1}\sqrt{\theta_z M_k \ln M_k} \tag{2-36}$$

其中，θ_z 的表达式由（2-31）给出。

引理 2.3. 首先，玩家 k 的阶段数目满足以下关系

$$2^{Z_k-1} \leq \frac{e-1}{\ln M_k} + \sqrt{\frac{(e-1)\hat{R}_{k,\max}}{M_k \ln M_k}} + \frac{1}{2} \tag{2-37}$$

因此，根据引理 2.2，可以得到

$$\begin{aligned} R_k &= \sum_{t=1}^{T} r_{k,\xi_t^k}(t) = \sum_{z=0}^{Z_k} \sum_{t=B_{k,z}}^{T_{k,z}} r_{k,\xi_t^k}(t) \\ &\geq \max_j \sum_{z=0}^{Z_k} \left(\sum_{t=B_{k,z}}^{T_{k,z}} \hat{r}_{k,j}(t) - 2\sqrt{e-1}\sqrt{\theta_z M_k \ln M_k} \right) \end{aligned} \tag{2-38}$$

利用关系式 $\theta_z = 4^z M_k \ln\left(M_k\right)/(e-1)$，上式可以简化为

$$\begin{aligned} R_k &\geq \max_{j \in \mathcal{A}_k} \hat{R}_{k,j}(T+1) - 2M_k \ln M_k \sum_{z=0}^{Z_k} 2^z \\ &= \hat{R}_{k,\max} - 2M_k \ln M_k \left(2^{Z_k+1} - 1\right) \end{aligned} \tag{2-39}$$

其次，根据引理 2.3，上式可以进一步转化为

$$\begin{aligned} R_k &\geq \hat{R}_{k,\max} + 2M_k \ln M_k - 8M_k \ln M_k \left(\frac{e-1}{\ln M_k} + \sqrt{\frac{(e-1)\hat{R}_{k,\max}}{M_k \ln M_k}} + \frac{1}{2} \right) \\ &= \hat{R}_{k,\max} - 2M_k \ln M_k - 8(e-1)M_k - 8\sqrt{e-1}\sqrt{\hat{R}_{k,\max} M_k \ln M_k} \end{aligned} \tag{2-40}$$

令 $f(x) = x - a\sqrt{x} - b\,(x \geq 0)$，其中，$a = 8\sqrt{e-1}\sqrt{M_k \ln M_k}$ 和常数项 $b = 2M_k \ln M_k + 8(e-1)M_k$。然后，对上式两边取对数，得到

$$\mathbb{E}[R_k] \geq \mathbb{E}[f(\hat{R}_{k,\max})] \tag{2-41}$$

由于函数 f 在定义域内可微，且其二阶导数为正数，因此 f 是一个凸函数。利用 Jensen 不等式，可以得到

$$\mathbb{E}[f(\hat{R}_{k,\max})] \geq f\left[\mathbb{E}[\hat{R}_{k,\max}]\right] \tag{2-42}$$

因此，当 $R_{k,\max} > a^2/4$，很容易获得 f 是在区间 $(a^2/4, +\infty)$ 的增函数，即 $f\left(\mathbb{E}[\hat{R}_{k,\max}]\right) \geq f\left(R_{k,\max}\right)$。结合（2-41）和（2-42），可以得到 $\mathbb{E}[R_k] \geq f\left(R_{k,\max}\right)$。当 $R_{k,\max} \leq a^2/4$，f 是在区间 $[0, a^2/4]$ 的减函数。因此，其最大值在 0 处取得，即 $f(0) = -b < 0 \leq \mathbb{E}[R_k]$。

最后，通过对网络中所有 FD 链路的遗憾求和，可以得到定理 2.2。 □

注 3. 定理 2.2 表明链路 k 的累积遗憾不会超过 $\mathcal{O}(\sqrt{R_{k,\max} M_k \ln M_k})$。由于 $R_{k,\max} \leq T$，因此，当总时隙 T 足够大时，链路每个回合的遗憾将趋于 0。也就是说，当总时隙 T 足够大时，SAO-FD-CSMA 算法将收敛。

2.5 仿真结果

本节通过数值仿真来验证所提算法的有效性。其中，网络参数设置主要参照 IEEE 802.11ax\g 标准，且所有结果均由 10^3 蒙特卡洛仿真得到。

2.5.1 参数设置与对比算法

首先，考虑信道指数路径衰减模型，即信道增益 $G_{\mathrm{Tx}_k,\mathrm{Rx}_k} = P_k D_0 d(\mathrm{Tx}_k, \mathrm{Rx}_k)^{-\alpha}$，其中，$\alpha$ 表示指数损耗因子，$d(\mathrm{Tx}_k, \mathrm{Rx}_k)$ 表示链路 k 发送端和接收端的欧式距离；此外，D_0 表示参考距离上的增益，即 $D_0 = (G_t G_r l^2)/((4\pi d_0)^2 L)$，其中，$G_t$ 和 G_r 分别表示天线的发射与接收增益；l 是中心频率 f_c 的波长，L 表示系统的硬件损耗，d_0 是参考距离；以上参数值设定为 $G_t = 1$，$G_r = 1$，$L = 1$ 和 $d_0 = 1\,\mathrm{m}$。网络中的背景噪声功率为 $-95\,\mathrm{dBm}$，FD 链路的自干扰消除因子 χ_k 为 $100\,\mathrm{dB}$。最后，网络中关于 CSMA 的参数由表 2.2 给出。

表 2.2 FD-CSMA 网络参数设置

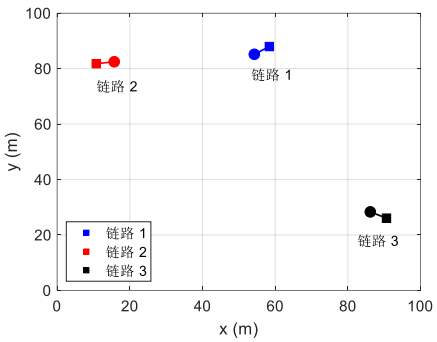
符号/缩写	物理意义	数值
$\mathbf{P}_k$	链路发送功率集合	$(10,15,20)$ dBm
$\mathbf{S}_k$	链路载波侦听门限集合	$(-70,-80,-90)$ dBm
$\mathbf{C}_k$	链路传输速率集合	$(20,50,100,150)$ Mbps
f_c	载波中心频率	5 GHz
B	信号带宽	40 MHz
t	单个时隙的长度	9 us
DIFS/SIFS	DISF 帧和 SIFS 帧长度	34/16 us
CW	CSMA 竞争窗口长度	32
L_{Data}	数据包长度	12000 比特
RTS/CTS	RTS 和 CTS 帧长度	160/120 比特
Epoch	数据帧长度	4×10^4 us
L_{ACK}	ACK 包长度	304 比特

在物理层的参数选择问题中，采用的对比算法包括：最优值（Optimal Value）方法，Exp3 算法、UCB1 算法、TS 算法，Tsallis-INF 算法和随机选择方法。其中，最优值方法通过穷尽搜索算法得到；Exp3 算法的探索和利用参数设置为0.04；UCB1 算法的估计的平均值的上界构造为$\hat{R}_{k,i}(t)/N_{k,i}(t)+\sqrt{2\log t/N_{k,i}(t)}$，其中，$N_{k,i}(t)$表示链路$k$直到时间$t$选择动作$i$的次数；对于 TS 算法，这里假设先验分布（即奖励过程）服从高斯分布，则其后验分布仍为高斯分布；Tsallis-INF 算法利用正则化 Tsallis 熵来代替 Exp3 算法中的动作概率分布的计算步骤；随机选择方法指链路每个时隙从其动作空间中随机选择一个动作进行传输。此外，由于以上对比算法都在算法 2.3 的框架下运行，所以下面将这些对比算法称为 Exp3-FD-CSMA 算法、UCB1-FD-CSMA 算法、TS-FD-CSMA 算法、Tsallis-INF-FD-CSMA 算法。

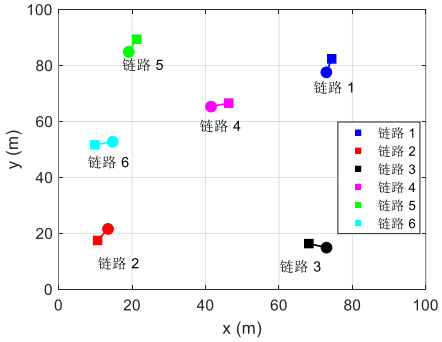
2.5.2 静态网络场景仿真

首先，考虑两种静态网络场景，如图 2.3a 和 2.3b 所示。在图 2.3a 中，有三条 FD 链路均匀的分布在一个(100×100) m 的区域内。其产生的过程为：首先随机产

生链路的发送端，然后接收端的角度和链路的长度分别随机地从区间$[0,\ 2\pi]$和$[d_{\min}, d_{\max}]$产生。与图 2.3a 不同的是，图 2.3b 中的链路数目由三条增加到六条。



（a）三条 FD 链路的静态网络场景



（b）六条 FD 链路的静态网络场景

图 2.3 两个静态网络场景图

图 2.4 给出了所提算法与对比算法在网络场景图 2.3a 中的网络吞吐量随时间变化的曲线，其中，部分参数设置为$\alpha=4$和$[d_{\min},d_{\max}]=[2,5]$，$T=12000$。从图中可以看出，除随机选择方法外，所有算法都可以收敛。具有先验信息（即动作权重$\omega_{k,i}$）的 SAO-FD-CSMA 算法的收敛速度最快，且拥有最好的性能；而 TS-FD-CSMA 算法的性能略低于所提算法，但其需要知道奖励的先验分布，这在实际系统中很难获得。此外，Tsallis-INF-FD-CSMA 算法的性能也接近于所提出的算法，但其需要反复调用牛顿方法，具有较高的时间复杂度。从图中还可以看到，UCB1-FD-CSMA 算法和 Exp3-FD-CSMA 算法的性能较差，这是因为 UCB1 算法常用于求解随机 MAB 问题，而 Exp3 算法常用于求解对抗 MAB 问题；但是，在该 FD-CSMA 网络中链路所处环境介于随机与对抗之间，所以两种算法的性能都不好。最后，与随机选择方法相比，所提算法的性能提高了大约 48%。

图 2.5 给出了所提算法与对比算法在网络场景图 2.3b 中的性能随时间变化的曲线，其参数设置与图 2.4 中相同。从图中可以看，对比三条链路的情况，网络的总吞吐量提高了，并且所有算法具有类似图 2.4 的增长趋势。然而，最优值和所提出算法之间的性能差距逐渐变大。其原因是：一方面，区域内的链路数越多，他们之间的竞争越激烈，从而导致更多的性能损失；另一方面，区域中的链路数越多，所提出的算法需要越多的时间收敛。

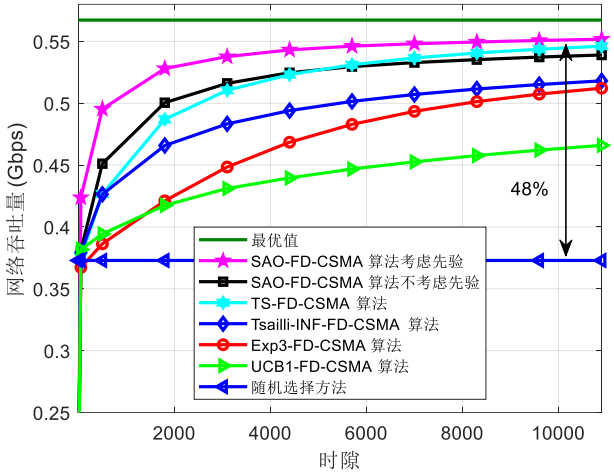


图 2.4 所提算法与对比算法在网络场景图 2.3a 中的性能随时间变化的曲线

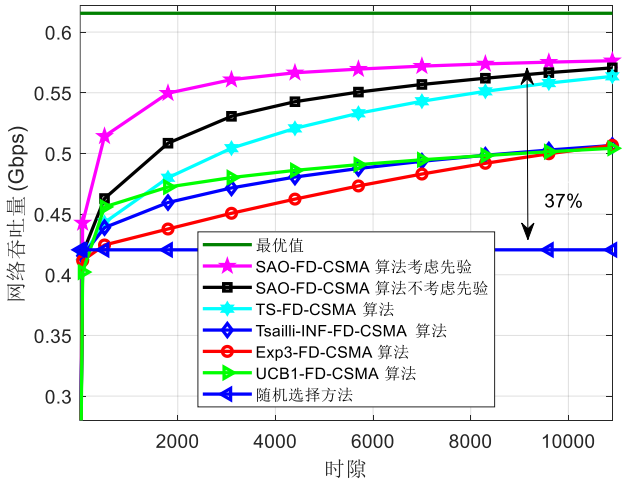


图 2.5 所提算法与对比算法在网络场景图 2.3b 中的性能随时间变化的曲线

2.5.3 动态网络场景仿真

动态网络场景是指：在每一次蒙特卡洛仿真中，FD 链路的位置在 (100×100) m 的区域内随机变化，链路长度在区间 $[d_{\min},d_{\max}]$ 内均匀变化。图 2.6 给出的是所提算法与对比算法在三条 FD 链路的动态网络下的性能曲线，其中，$\alpha=4$ 和 $[d_{\min},d_{\max}]=[2,5]$。从图中可以看到所有算法的网络吞吐量低于图 2.4 中的结果，但都具有相同的增长趋势。而且，与图 2.4 不同的是，这里有先验信息的 SAO-FD-CSMA 算法比没有先验信息的 SAO-FD-CSMA 算法稍差。这是因为算法无法获得所有 10^3 次蒙特卡洛仿真下动态网络的先验信息。从图 2.6 还可以看出，TS-FD-

CSMA 算法与没有先验信息的 SAO-FD-CSMA 算法具有相似的性能，但 TS-FD-CSMA 算法需要一些知道奖励分布这一先验信息。此外，与随机选择方法相比，所提算法的性能提高了约 43%。

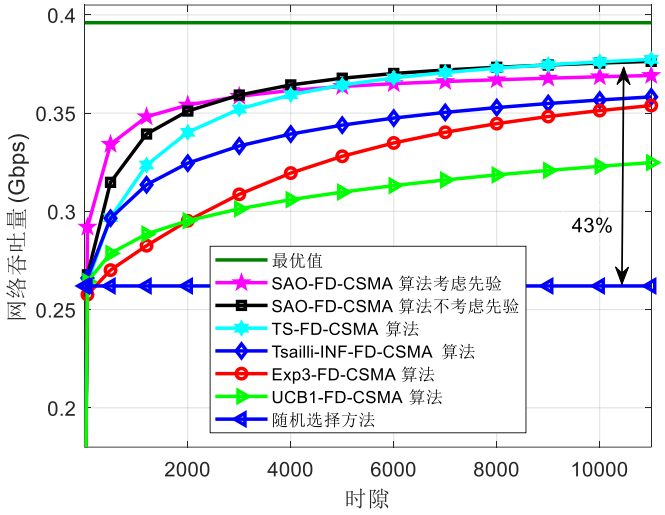


图 2.6 所提算法与对比算法在10^3次动态网络场景下的性能随时隙变化的曲线

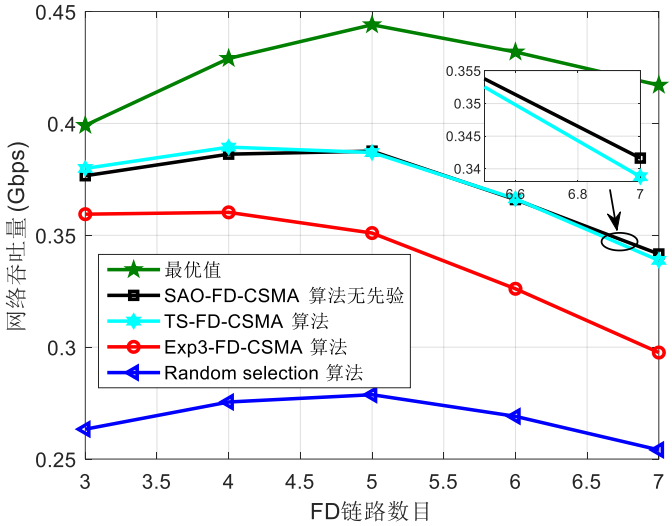


图 2.7 所提算法与对比算法在10^3次动态网络场景下性能随链路数目变化的曲线

图 2.7 描绘了所提算法、最优方法、TS-FD-CSMA 算法、Exp3-FD-CSMA 算法和随机选择方法在10^3动态网络场景下网络吞吐量随链路数目变化的曲线，其中，$\alpha=4$，$[d_{\min},d_{\max}]=[2,5]$，$T=12000$。从图中可以看到，除 Exp3-FD-CSMA 算法外，其他算法在链路数目数从 3 增加到 5 时先增加，然后在链路数超过 5 时快速下

降。而且，TS-FD-CSMA 算法在链路数小于5时性能最佳，但当链路数大于6时，它的性能略低于所提算法。这是因为随着网络中的链路数的增多，TS-FD-CSMA 算法需要遍历所有网络状态所需的时间也就越多。此外，由于链路数目增大，网络中链路之间的竞争增加，使得最优方法与其他算法之间的性能差距越来越大。因此，在高密度网络中可以选择较大的T，而在低密度网络中选择较小的T。

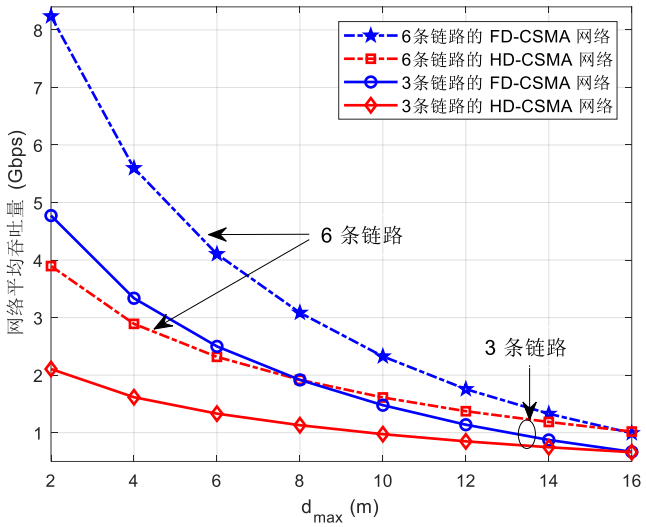


图 2.8 所提算法在 HD-CSMA 网络和 FD-CSMA 网络的性能随着 FD 链路长度 $d_{\max}$ 变化的性能曲线

最后，比较链路长度对 HD-CSMA 网络和 FD-CSMA 网络的性能的影响。图 2.8 给出的是所提算法在10^3动态网络场景下网络平均吞吐量随链路长度变化的曲线，其中，平均吞吐量可以表示为$\sum_{k=1}^{K}\sum_{s=1}^{T}\Gamma_k(s)/T$，其中$T=1.2\times10^4$。为了更好地捕捉 FD-CSMA 网络和 HD-CSMA 网络之间的差异性，仿真中假设链路的传输速率是连续的。从图中可以看到，FD-CSMA 网络和 HD-CSMA 网络的平均吞吐量都随着链路长度增加而下降。这是因为增加$d_{\max}$，会降低接收信号功率，同时放大 SINR 中的自干扰。与 HD-CSMA 网络相比，FD-CSMA 网络在 3 链路和 6 链路情况下都遭受更严重的性能下降。值得注意的是，当$d_{\max}>16(\mathrm{m})$时，3 链路和 6 链路下的 FD-CSMA 网络的性能都低于 HD-CSMA 网络。

2.6 本章小结

本章研究了 FD-CSMA 网络的空间复用问题，联合考虑物理层的 TP 和 CST 选择和 MAC 层的 LAI 参数优化，来最大化网络的平均吞吐量。首先，基于 FD 信号模型、CSMA 载波侦听模型和 CTMN 模型，将该问题建模为一个传统的网络优化问题。然后，利用分解理论将其分解为两个子问题。针对子问题一，提出了一种最优 FD-CSMA 算法，通过次梯度下降方法找到每条链路最优的 LAI；针对子问题二，提出了一种 SAO 算法，通过序贯决策方式为每条链路选择最优的 TP 和 CST 组合。最后，通过交替迭代求解子问题一和子问题二，进一步提出了 SAO-FD-CSMA 算法来近似解决该网络吞吐量最大化问题。为了评估所提算法的有效性，设计了一种基于 Matlab 的 DESim 仿真平台来模拟 CSMA 协议。基于该平台，通过数值仿真结果验证了所提算法的有效性，与随机选择方法相比，所提算法在静态网络和动态网络场景下分别提升了约 48% 和 43%。此外，还理论分析了所提算法的收敛性和遗憾上界，即当时间 $T \to \infty$ 时，SAO-FD-CSMA 算法可以收敛到最优值。

第 3 章 基于 MPMAB 的分布式异构网络资源分配策略

本章介绍基于 MPMAB 的分布式异构网络资源分配策略[①]。首先，介绍异构网络中 IoT 设备的两种传输模式和最优化问题模型；其次，将联合智能反射面和扩频因子分配问题建模成一个两阶段的 MPMAB 问题；接着，利用非合作博弈理论和 MAB 技术，提出一种 E2Boost 算法来求解该两阶段 MPMAB 问题；最后，推导所提算法的累积遗憾上界，且通过数值仿真验证所提算法的有效性。

3.1 引言

可重构智能反射面（Reconfigurable Intelligent Surface, RIS）通过控制二维平面上的大量低成本的无源二极管的通断，调整入射信号的幅度和相位等参数，重新构建适合于信号传播的信道，大幅提高链路通信质量，得到越来越多的关注[107]。其中，RIS 辅助的蜂窝物联网（Cellular Internet-of-Things，C-IoT）因低成本、大规模和超连接的能力，被认为是下一代无线通信网络的范例之一[108]。通过采用 LoRa（Long Range）技术，C-IoT 可以在未经授权的频段上工作。然而，长距离通信和工作在未授权频段的特点，使得 C-IoT 信号对周边的干扰非常敏感。为了提高抗干扰能力，C-IoT 在物理层采用不同扩频因子（Spreading Factor, SF）对传输信号进行扩频调制，实现速率自适应的同时提高抗干扰能力[②]。目前，关于 C-IoT 在多 RIS 辅助的异构蜂窝网中的研究仍处于初步阶段。

本章考虑异构蜂窝网络中的上行通信链路，在该异构网络中存在蜂窝网用户和 IoT 设备。其中，IoT 设备需要将数据传输至基站（Base Station, BS），由于障碍物的存在，IoT 设备到基站的信道可能经历深度衰落，它需要机会性地接入 RIS 辅助的蜂窝网络、并选择合适的 SF 来提高其传输速率。系统目标是通过为每个 IoT 设备寻找最佳的 RIS 和 SF 来最大化网络中所有 IoT 设备的传输速率之和。该速率

[①] 本章内容已发表于 IEEE Transactions on Communications.

[②] 由于 LoRa 技术采用扩频调制方式，因此 CIoT 的传输信号即使淹没在噪声或干扰中，接收端仍能正常解调有用信号。

最大化问题面临以下挑战：第一，没有中心节点来协调 IoT 设备之间的传输，即 IoT 设备之间无法共享收集到的信道状态信息（Channel State Information, CSI）和 RIS 的状态信息；第二，在该异构蜂窝网络中，RIS 是部署给蜂窝网用户的，IoT 设备没有任何 RIS 的相位和幅度等信息。

为了克服以上挑战，本文采用在线学习方法将该联合 RIS 与 SF 选择问题建模成一个两阶段的 MPMAB 框架。其中，玩家是 IoT 设备，奖励是传输成功或失败的反馈，第一阶段 MPMAB 的动作是 RIS，第二阶段 MPMAB 的动作是 SF。通过结合通信理论、博弈理论和 MAB 技术，提出一种 E2Boost（Exploration and Exploitation Boosting）算法来求解该两阶段 MPMAB 问题。所提算法又分成三个阶段：ϵ-Greedy EE 阶段、非合作博弈阶段和 TS（Thompson sampling）EE 阶段。因为每个阶段都包含一个特定的机制来权衡 EE 困境，所以将该算法称为 E2Boost 算法。利用有限时间分析方法，推导了 E2Boost 算法的累积遗憾上界，表明所提算法的性能比已有的分布式分配策略提高了约 M 倍，其中 M 是物联网设备的扩频因子数目。最后，利用仿真结果验证了所提算法的有效性。

3.2 系统模型

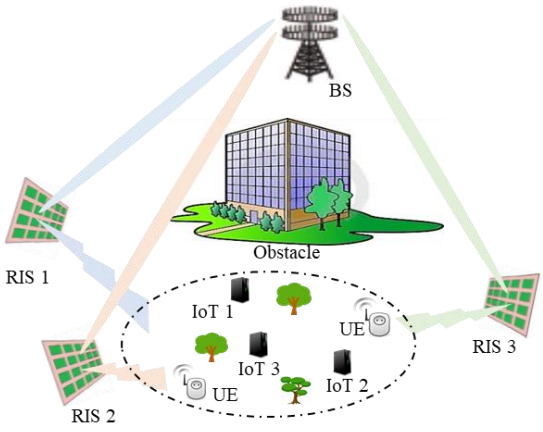


图 3.1 RIS 辅助的异构蜂窝网示意图

考虑一个上行 RIS 辅助的异构蜂窝网络，如图 3.1 所示。其中，几个蜂窝网用户（User，UE）和 N 个 IoT 设备位于同一个区域内。UE 和 IoT 设备都需要向 BS 传输数据。由于 UE 和 BS 之间存在一些障碍物（比如建筑、树木等），信号

可能经历深度衰落。因此，K 个 RIS 被部署在该区域内辅助 UE 将信号传输到 BS。假设 RIS 工作在不同的频率上，且 IoT 设备没有任何关于这些 RIS 的信息。为了提高通信速率，IoT 设备需要机会性地接入空闲的 RIS 以提高其传输速率。为防止其对 UE 用户的通信产生干扰，IoT 设备需要先执行频谱感知操作。假设时间离散化为$t=1,2,\ldots,T$，且在每个时隙中一个 RIS 可以服务多个 UE，但只能被一个 IoT 设备占用。系统目标是通过为每个 IoT 设备寻找最佳的 RIS 和 SF 来最大化网络中所有 IoT 设备的传输速率之和。

3.2.1 信道模型

在该异构网络中，每个 IoT 设备存在两种传输模式。第一，RIS 辅助的传输模式，即当检测到目标 RIS 处于空闲状态时，IoT 设备接入该 RIS 上传数据至 BS；第二，直传模式，即当检测到目标 RIS 处于忙碌状态，IoT 设备则直接以较低的速率传输至 BS。

对于模式一，假设 RIS 上的每一个元素有b个二极管。通过控制这些二极管的通断，总共可以在$[0,\ 2\pi)$区间内产生2^b个相位。因此，RIS 上第l_1行、l_2列元素的相位可以表示为

$$\tau_{l_1,l_2}=\frac{\pi\rho_{l_1,l_2}}{2^{b-1}} \tag{3-1}$$

其中，ρ_{l_1,l_2}是在区间$[0,2^b-1]$内的一个整数。令A_{l_1,l_2}表示 RIS 上第l_1行、l_2列元素的反射系数，即

$$A_{l_1,l_2}=Ae^{-j\tau_{l_1,l_2}} \tag{3-2}$$

其中，A表示在区间$(0,\ 1]$内的一个常数。

由于 RIS 的辅助，IoT-RIS-BS 链路的信号通常强于其它路径（包括 IoT-BS 直射路径）的信号。因此，IoT 设备与 BS 之间的信道可以用莱斯模型描述，即

$$h_{l_1,l_2}^{n,k}=\sqrt{\frac{\zeta}{\zeta+1}}\tilde{h}_{l_1,l_2}^{n,k}+\sqrt{\frac{1}{\zeta+1}}\hat{h}_{l_1,l_2}^{n,k} \tag{3-3}$$

其中，$\tilde{h}_{l_1,l_2}^{n,k}$和$\hat{h}_{l_1,l_2}^{n,k}$分别表示第$k$个 RIS 与第$n$个 IoT 设备经过 RIS 第$(l_1,l_2)$个元素的直视分量和非直视分量；符号$\zeta$表示莱斯因子，即直视分量与非直视分量之比。为

了简便，下面省略 IoT 设备n与 RISk的符号下标。令D_{l_1,l_2}表示 BS 与 RIS 第(l_1,l_2)个元素之间的距离，以及d_{l_1,l_2}表示 IoT 设备与 RIS 第(l_1,l_2)个元素之间的距离，则 IoT 设备传输至 BS 并经过 RIS 第(l_1,l_2)个元素的距离为$L_{l_1,l_2}=D_{l_1,l_2}+d_{l_1,l_2}$。根据文献[114]，IoT 设备至 BS 经过 RIS 第(l_1,l_2)个元素的直视分量为

$$\begin{aligned}\tilde{h}_{l_1,l_2}&=\sqrt{GD_{l_1,l_2}^{-\iota}d_{l_1,l_2}^{-\iota}}e^{-j\frac{2\pi}{\lambda}L_{l_1,l_2}}\\&=\sqrt{G}\left[\sqrt{D_{l_1,l_2}^{-\iota}}e^{-j\frac{2\pi}{\lambda}D_{l_1,l_2}}\right]\left[\sqrt{d_{l_1,l_2}^{-\iota}}e^{-j\frac{2\pi}{\lambda}d_{l_1,l_2}}\right]\end{aligned}\tag{3-4}$$

其中，ι表示路径指数损耗因子；G表示天线增益；λ表示信号波长。此外，路径上的非直视分量为

$$\hat{h}_{l_1,l_2}=\sqrt{PL_{\mathrm{NLoS}}(L_{l_1,l_2})}g_{l_1,l_2}\tag{3-5}$$

其中，g_{l_1,l_2}表示小尺度非直视分量，服从独立同分布的复高斯分布，即$g_{l_1,l_2}\sim\mathcal{CN}(0,1)$；符号$PL_{\mathrm{NLoS}}(\cdot)$表示非直视分量的信道增益。

对于模式二，IoT 设备直接向 BS 传输数据，无需 RIS 的辅助。因为 IoT 设备与基站之间存在一些障碍，信号可能经历深度衰落。因此，可以利用阴影衰落信道模型对此模式下的 IoT 设备与 BS 之间的信道进行描述，即

$$h_n=\sqrt{\varrho_n}g_n\tag{3-6}$$

其中，ϱ_n表示信道的功率增益，服从均值为μ_n、标准方差为σ_n的独立同分布的对数正态分布。在实际无线信道中，σ_n的典型值在6和12 dB 之间[109]。另外，g_{l_1,l_2}表示小尺度非直视分量，服从独立同分布的复高斯分布，即$g_{l_1,l_2}\sim\mathcal{CN}(0,1)$；符号$PL_{\mathrm{NLoS}}(\cdot)$表示非直视分量的信道增益。

3.2.2 信号模型

经过信道传输后，BS 接收到第n个 IoT 设备的信号可以表示为

$$\begin{cases}F_{n,k}=\sum\limits_{l_1,l_2}A_{l_1,l_2}^{k}h_{l_1,l_2}^{n,k}\sqrt{\Omega_n}x+y+\omega, & \text{模式一}\\ F_n=h_n\sqrt{\Omega_n}x+y+\omega, & \text{模式二}\end{cases}\tag{3-7}$$

其中，x 表示传输信号，且功率归一化为1；y 表示接收到的干扰信号，服从对数正态分布[110]，即 $y \sim Log\mathcal{N}(\mu_y, \sigma_y^2)$。此外，$\omega$ 表示独立同分布的复高斯信号；Ω_n 表示发送功率。因此，接收到的信干噪比（Signal-to-Interference-plus-Noise Ratio, SINR）可以表示为

$$\begin{cases} \gamma_{n,k} = \dfrac{\Omega_n \left(\sum\limits_{l_1,l_2} A_{l_1,l_2}^k \tilde{h}_{l_1,l_2}^{n,k} \sum\limits_{l_1,l_2} \left(A_{l_1,l_2}^k \right)^* \left(\tilde{h}_{l_1,l_2}^{n,k} \right)^* \right)}{\exp\left(2\mu_y + 2\sigma_y^2\right) + \sigma_\omega^2}, & \text{模式一} \\ \gamma_n = \dfrac{\Omega_n h_n h_n^*}{\exp\left(2\mu_y + 2\sigma_y^2\right) + \sigma_\omega^2}, & \text{模式二} \end{cases} \tag{3-8}$$

其中，$(\cdot)^*$ 表示共轭运算。在实际系统中，IoT 设备通常仅能支撑有限多个传输速率，即 SF 的数目。令 $\mathcal{M} = \{c_1, c_2, \cdots, c_M\}$ 和 $\mathcal{S} = \{s_1, s_2, \cdots, s_M\}$ 分别表示传输速率和 SF。根据文献[111]，传输速率与 SF 之间存在以下关系

$$c_m = \frac{Bs_m}{2^{s_m}} \times CR \tag{3-9}$$

其中，B 表示带宽；CR 表示编码速率。从上式可以看到，SF 越大设备传输的速率越小。

然而，设备的传输速率不仅取决于编码方式，还和当前的接收 SINR 有关。因此，IoT 设备在速率 c_m 上的成功传输概率可以定义为

$$\begin{cases} \theta_{k,c_m}^n \triangleq \Pr\{\gamma'_{n,k} \geq \Psi_m\}, & \text{模式一} \\ \theta_{c_m}^n \triangleq \Pr\{\gamma'_n \geq \Psi_m\}, & \text{模式二} \end{cases} \tag{3-10}$$

其中，Ψ_m 表示解调当前信号所需的最小接收 SINR。值得注意的是，由于信道的小尺度非直视分量和干扰信号，瞬时接收 SINR $\gamma'_{n,k}$ 可以看作是一个均值为 $\gamma_{n,k}$ 的随机变量。根据香农公式，对于给定的传输速率，接收 SINR 越大，信号的传输成功概率越高。因此，一个降序排列的传输速率 $\left(c_1 > c_2 > \cdots > c_M\right)$ 对应于一个升序排列的传输成功概率 $\left(\theta_{c_1} < \theta_{c_2} < \cdots < \theta_{c_M}\right)$。

3.2.3 最优化问题模型

系统目标是为每个 IoT 设备寻找最佳的传输参数来最大化异构网络中所有 IoT 设备的传输速率之和。具体地，在模式一下寻找最佳的 RIS 与 SF 组合；在模式二下寻找最佳的 SF。令 $\vec{\vartheta}^t=\{\vartheta_1^t,\vartheta_2^t,\ldots,\vartheta_K^t\}$ 表示 RIS 在时隙 t 的状态向量（空闲或占用），其中，$\vartheta_k^t=1$ 表示第 k 个 RIS 处于空闲状态；否则，其被 UE 占用。假设 IoT 设备利用频谱感知技术预先知道 RIS 的状态信息。则该资源分配问题可以表述为

$$\begin{aligned}
\max_{\phi_{k,c_m}^n,\psi_{c_m}^n}\quad & \sum_{t=1}^{T}\left(\underbrace{\sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{m=1}^{M}c_m\vartheta_k^t\theta_{k,c_m}^n\phi_{k,c_m}^n}_{\text{模式一}}+\underbrace{\sum_{n=1}^{N}\sum_{m=1}^{M}c_m\theta_{c_m}^n\psi_{c_m}^n}_{\text{模式二}}\right)\\
\text{s.t.}\quad & \sum_{m=1}^{M}\sum_{k=1}^{K}\vartheta_k^t\phi_{k,c_m}^n+\sum_{m=1}^{M}\psi_{c_m}^n=1,\ \forall n\in\mathcal{N} \qquad (3\text{-}11)\\
& \sum_{n=1}^{N}\phi_{k,c_m}^n\le 1,\ \forall c_m\in\mathcal{M},\forall k\in\mathcal{K}
\end{aligned}$$

其中，ϕ_{k,c_m}^n 和 $\psi_{c_m}^n$ 是两个二元变量，且 $\phi_{k,c_m}^n=1$ 表示 IoT 设备采用 SF m 经 RIS k 将数据传输至 BS；否则，$\phi_{k,c_m}^n=0$。同时，$\psi_{c_m}^n=1$ 表示 IoT 设备采用 SF m 直接将数据传输至 BS；否则，$\psi_{c_m}^n=0$。上式第一个约束表明每个 IoT 设备要么传输在模式一，要么在模式二。如果 IoT 设备 n 传输在模式一，则 $\sum_{m=1}^{M}\sum_{k=1}^{K}\phi_{k,c_m}^n=1$ 表示每个 IoT 设备应该至少选择一个 RIS 和 SF 进行传输；如果其传输在模式二，则 $\sum_{m=1}^{M}\psi_{c_m}^n=1$ 表示每个 IoT 设备应该选择一个 SF 进行传输。第二个约束表明选择第 k 个 RIS 和 m 个 SF 的 IoT 设备的数量最多为 1。此外，$\mathcal{N}=\{1,2,\cdots,N\}$ 和 $\mathcal{K}=\{1,2,\cdots,K\}$ 分别表示 IoT 设备和 RIS 的集合。符号 θ_{k,c_m}^n 表示 IoT 设备 n 在第 k 个 RIS 和 m 个 SF 上的成功传输概率；而 $\theta_{c_m}^n$ 表示 IoT 设备 n 在第 m 个 SF 上的成功传输概率。

在分布式异构网络中，直接求解问题（3-11）是困难的。首先，c_m 和 θ_{k,c_m}^n 是离散值，使得目标函数与约束条件是非凸的；其次，直接求解（3-11）需要知道 θ_{k,c_m}^n 的准确值，因为信道特性是由 UE 控制的 RIS 决定的，这个信息很难获得；最后，IoT 设备之间需要一些信息传递确定每个 IoT 设备的最佳 RIS，但在该分布式网络

中是不可行的。为了克服这些挑战，可以利用在线学习方法来学习 $\theta_{c_m}^n$ 和 θ_{k,c_m}^n 的值，并自适应地将为每个 IoT 设备分配最佳的 RIS 和 SF。在此过程中，IoT 设备不仅需要充分地探索 RIS 和 SF 的组合，还需要在每个时隙尽可能的利用当前最好的 RIS 和 SF 进行数据传输。为了有效权衡这个 EE 困境，本章引入了 MPMAB 框架来解决这个问题，其中玩家是 IoT 设备，奖励是传输成功或失败的反馈，动作在模式一中为 RIS，在模式二中为 SF。

然而，由于动作空间（即 RIS 和 SF 的组合）可能较大，MPMAB 框架面临收敛速度慢的问题。为了克服这个问题，下一节将该 MPMAB 问题解耦为一个两阶段的 MPMAB 问题来序贯地学习出最优的 RIS 和 SF，从而提高算法的收敛速度。该问题可以解耦的原因是：一方面，按升序排列的传输速率将导致降序排列的成功传输概率；另一方面，具有不同传输速率的 IoT 设备在特定 RIS 下经历相同的信道衰落。因此，IoT 设备传输速率在不同 RIS 上将有相同趋势的平均成功传输概率，即可以通过为 IoT 设备分配任意速率来估计 RIS 的平均成功传输概率，从而确定该设备的最佳 RIS。换句话说，SF 与 RIS 的分配过程在模式一下是相互独立的。

3.2.4 MPMAB 问题模型

（1）第一阶段 MPMAB 问题

在这个阶段，动作是 RIS，奖励是传输成功或失败的反馈，目标是为每个 IoT 设备分配最佳的 RIS。令 $X_{I'_{n,t}}(t)$ 表示 IoT 设备 n 传输在动作 $I'_{n,t}$ 上的反馈，则 $X_{I'_{n,t}}(t)=1$ 表示传输成功；否则，$X_{I'_{n,t}}(t)=0$。在该分布式网络，当多个 IoT 设备同时选择同一个 RIS 时会发生碰撞。令 η 表示碰撞指示，当 IoT 设备没有从 BS 收到任何反馈时，则认为碰撞发生，这时 $\eta=0$；否则，$\eta=1$。定义符号 $\mathbf{I}'_t=\{I'_{1,t},I'_{2,t},t\cdots,I'_{N,t}\}$ 网络中 IoT 设备的选择策略，则第 k 个 RIS 是否发生碰撞可以表示为

$$\eta_k\left(\mathbf{I}'_t\right)=\begin{cases}0, & |\mathcal{N}_k|>1,\\ 1, & \text{其它},\end{cases} \tag{3-12}$$

其中，$\mathcal{N}_k$ 表示在策略 $\mathbf{I}'_t$ 中选择第 k 个 RIS 的 IoT 设备的集合。因此，IoT 设备 n 选

择第 k 个 RIS 的奖励和传输成功概率分别为

$$r_{n,I'_{n,t}=k}(t) \triangleq \eta_k\left(\mathbf{I}'_t\right) X_{I'_{n,t}=k}(t) \tag{3-13}$$

和

$$\hat{\theta}_{n,k} = \mathbb{E}\left[r_{n,I'_{n,t}=k}(t)\right] \tag{3-14}$$

其中，$\mathbb{E}[\cdot]$表示期望运算。

（2）第二阶段 MPMAB 问题

在这个阶段，动作是 SF，奖励仍然是传输成功与失败的反馈，目标是为每个 IoT 设备分配最佳的 SF。需要注意的是，这个阶段需要同时处理两种传输模式。在模式一下，IoT 设备基于第一段的最佳 RIS 来寻找最佳的 SF；而在模式二下，IoT 设备直接寻找最佳的 SF。由于 SF 的选择不存在碰撞问题，因此，第二阶段 MPMAB 问题也可以看作是单个玩家的 MAB 问题。同样，令 $I^{''}_{n,t}$ 表示当前时隙被选中的动作，则 IoT 设备 n 利用第 m 个速率进行传输的奖励为

$$r_{n,I''_{n,t}=m}(t) \triangleq c_m \eta_k\left(\mathbf{I}'_t\right) X_{I''_{n,t}=m}(t) \tag{3-15}$$

其中，$\mathbf{I}'_t$ 表示在第一阶段所有 IoT 设备的关于 RIS 的选择策略。因此，选择第 m 个 SF 的估计的平均奖励为

$$\hat{\mu}_{n,m} = \mathbb{E}\left[r_{n,I''_{n,t}=m}(t)\right] = c_m \hat{\theta}_{n,m} \tag{3-16}$$

其中，$\hat{\theta}_{n,m}$ 表示 IoT 设备利用第 m 个速率进行传输的估计的成功概率。

3.3 智能反射面与扩频因子联合分配策略

3.3.1 最优智能反射面分配策略

最优 RIS 分配策略的核心思想是利用 IoT 设备在各 RIS 上估计的成功传输概率作为效用函数，设备之间通过非合作博弈来得到各自最佳的 RIS，其伪代码由算法 3.1 给出。首先，IoT 设备需要学习出各 RIS 上的成功传输概率。一种方法是 IoT 设备对每个 RIS 进行多次尝试，然后根据其传输成功或失败的反馈计算传输成功概率。但是，当 RIS 数目较多时，这种方式需要很难收敛到真实的成功传输概率。

因此，本节提出一种基于ε-贪婪算法的 EE 选择策略来获取估计的传输成功概率，即 IoT 设备以ε的概率均匀地探索所有 RIS，以1−ε的概率探索上一阶段最好的 RIS。

算法 3.1 最优 RIS 分配策略（在 IoT 设备n上执行）

步骤 1： 初始化参数：$\delta>0,\varepsilon>0,\nu_1,\nu_2>0$；同时设定$V_{n,k}(0)=0,Q_{n,k}(0)=0,\epsilon=1,\forall k\in\mathcal{K},\forall m\in\mathcal{M},ST_n=C$

ϵ-贪婪算法的 EE 阶段：

步骤 2: 执行$\nu_1 z^{\delta}$次步骤 3 至步骤 7

步骤 3: 从集合中$\mathcal{M}$中随机选择一个传输速率，即$I'_{n,t}=m$

步骤 4: 以ϵ的概率从集合中$\mathcal{K}$中选择一个 RIS $I'_{n,t}=k$；否则，以$1-\epsilon$的概率选择当前最优的 RIS $I'_{n,t}=k_n^*$

步骤 5: 利用频谱感知技术确定当前 RIS 的状态：若占用，直接进行最优 SF 分配；否则，继续执行以下步骤

步骤 6： 观测传输反馈$X_{n,I'_{n,t}}$，并且记录碰撞指示η_k

步骤 7： 利用公式（3-16）更新动作I_a^t的经验平均值$\hat{\mu}_{I_a^t}(t)$

步骤 8： 更新所有 RIS 的估计的成功概率：$\hat{\theta}_{n,k}^z=Q_{n,k}/V_{n,k}$， $\forall k\in\mathcal{K}$

非合作博弈阶段：

步骤 9： 执行$\nu_2 z^{\delta}$次步骤 10 至步骤 12

步骤 10： 若$ST_n=C$，则根据式（3-18）来选择 RIS；如果$ST_n=D$，则根据式（3-19）来选择 RIS

步骤 11： 利用频谱感知技术确定当前 RIS 的状态：若占用，直接进行最优 SF 分配；否则，继续执行以下步骤

步骤 12： 根据式（3-20）和（3-21）来更新当前状态，即状态C与D的转换

步骤 13： 根据引理 3.1 更新参数ϵ

步骤 14： 确定当前最优的 RIS：$\arg\max_{k\in\mathcal{K}}\sum_{j=0}^{z/2}F_n^{z-j}(k)$

其次，基于估计的传输成功概率，IoT 设备之间进行非合作博弈。具体地，令

$\hat{\theta}_{n,k}^{z}$ 表示 IoT 设备 n 在 z 个时间段内选择第 k 个 RIS 的估计的平均传输概率，则 IoT 设备 n 的效用函数可以表示为

$$u_n(\mathbf{I}') \triangleq \eta_k(\mathbf{I}')\hat{\theta}_{n,k}^{z}, \forall k \in \mathcal{K} \tag{3-17}$$

其中，$\mathbf{I}'$ 表示网络中 IoT 设备关于 RIS 的选择策略。假设每个 IoT 设备有一个私有状态 $ST_n = \{C, D\}, \forall n \in \mathcal{N}$，其中 C 表示同意状态，D 表示不同意状态。同时，假设每个 IoT 设备有一个基准 RIS $\bar{k}$，且令 $u_{n,\max} = \max_{\mathbf{I}'} u_n(\mathbf{I}')$，则 IoT 设备间的非合作博弈策略可以定义为：

· 一个处于状态 C 的 IoT 设备很大概率在下一回合选择基准 RIS $\bar{k}$，即

$$P_{n,k} = \begin{cases} \dfrac{\varepsilon^{\nu}}{K-1}, & k \neq \bar{k}; \\ 1-\varepsilon^{\nu}, & k = \bar{k}. \end{cases} \tag{3-18}$$

· 一个处于状态 D 的 IoT 设备在下一回从集合 $\mathcal{K}$ 中随机选择一个 RIS，即

$$P_{n,k} = \frac{1}{K}, \forall k \in \mathcal{K} \tag{3-19}$$

此外，状态 C 和 D 的转换满足以下关系：

· 若 $k = \bar{k}, u_n > 0, ST_n = C$，则处于状态 C 的 IoT 设备仍为状态 C 的概率为1，

$$(\bar{k}, C) \to (\bar{k}, C) \tag{3-20}$$

· 若 $k \neq \bar{k}$ 或 $u_n = 0$ 或 $ST_n = D$，则基准 RIS $\bar{k}$ 和状态的转移可以表示为

$$\left(\bar{k}, C / D\right) = \begin{cases} \left(k, C\right), & \dfrac{u_n}{u_{n,\max}} \varepsilon^{u_{n,\max} - u_n}; \\ \left(k, D\right), & 1 - \dfrac{u_n}{u_{n,\max}} \varepsilon^{u_{n,\max} - u_n}. \end{cases} \tag{3-21}$$

上式表明，当 RIS 观察到碰撞或处于忙碌状态时，由于 $u_n = 0$，IoT 设备将以100%的概率转移到状态 D。另一方面，若当前 RIS 是 IoT 设备的最优动作，则其它以100%的概率转移到状态 C。

假设所有玩家的动作与状态构成策略 $\mathbf{a_1}$，则在博弈阶段结束后将得到一个策略图。图中的顶点是每个策略，两顶点之间是否存在边取决于玩家能否从该顶点策

略切换到另一顶点策略。实际上，这个策略图在状态空间 $\prod_{n=1}^{N}\left(\mathcal{K}_n \times (C,D)\right)$ 中构成了一个扰动的时间可逆马尔可夫过程。根据文献[112]，可以证明图中存在唯一一个最佳的策略，使得玩家访问该最佳策略的次数明显高于其他策略。因此，每个玩家只需记录各个动作被选择的次数便可以在完全分布式的情况下确定最优动作。具体地，玩家 n 直至时间段 z 选择第 k 个 RIS 的次数可以表示为

$$F_n^z(k) \triangleq \sum_{t\in\mathcal{G}_z} \mathbb{I}\left\{I'_{n,t}=k, ST_n=C\right\}, \forall k \in \mathcal{K} \tag{3-22}$$

其中，$\mathcal{G}_z$ 表示时间段 z 内的总时隙数目，$\mathbb{I}\{\cdot\}$ 表示指示函数。最后，当前时间段内最佳的 RIS 可以确定为

$$k_n^* = \arg\max_{k\in\mathcal{K}} \sum_{j=0}^{z/2} F_n^{z-j}(k) \tag{3-23}$$

下面给出引理 3.1 来自适应地调节 EE 参数 ϵ，加速 ϵ-贪婪算法的收敛速度。

引理 3.1. 当 $z>1$ 时，对于玩家 n，EE 参数 ϵ 可以通过下式进行调整

$$\epsilon \triangleq \min\{1, \mathrm{D}_{\mathrm{WD}}\left(\mathbb{P}(F_n^z) \| \mathbb{P}(F_n^{z-1})\right)\} \tag{3-24}$$

其中，$\mathrm{D}_{\mathrm{WD}}\left(\cdot\|\cdot\right)$ 表示 WD（Wasserstein Distance）距离的运算符号；同时，F_n^z 和 F_n^{z-1} 分别表示时间段 z 和 $z-1$ 内所有 RIS 被选中次数的概率质量函数，即 $\mathbb{P}(F_n^z(i)) = F_n^z(i) / \sum_i F_n^z(i), \forall i \in \mathcal{K}$。

证明：当非合作博弈阶段的解接近最优 RIS 时，算法没有必要再均匀地探索所有的 RIS。很显然，这种渐近行为可以通过两个相邻向量 F_n^z 和 F_n^{z-1} 之间的距离来刻画。本节采用 WD 距离来衡量两向量的概率质量函数的距离。接着，利用该 WD 距离来调节 EE 参数 ϵ。然后，利用调整后的 ϵ，运行算法 3.1，可以获得新的向量 F_n^z 和 F_n^{z-1}。可以观察到，他们之间的距离将逐渐缩小，即 ϵ 逐渐变小；所以，算法 3.1 将以很大的概率（即 $1-\varepsilon$）选择当前最佳的 RIS。 □

3.3.2 最优扩频因子分配策略

根据之前的分析，SF 分配可以看成是单个玩家的 MAB 问题。此外，奖励是传输成功或失败的二元反馈，可以看成是贝努利分布，而动作上的成功传输概率又可以看成是 Beta 分布。因此，本节采用 TS 算法来求解该 MAB 问题。需要注意的是这个阶段需要处理两种传输模式。在模式一，IoT 设备需要基于上一阶段的最佳 RIS 将数据传输至 BS；在模式二，IoT 设备不需要 RIS 辅助而直接传输至 BS。下面给出两种模式下 SF 分配策略的伪代码，如算法 3.2 所示。

算法 3.2 最优 SF 分配策略（在 IoT 设备 n 上执行）

步骤 1： 初始化参数：$\delta>0,\nu_3>0$；同时设定 $\alpha_{n,m}(0)=0,\beta_{n,m}(0)=0,\forall m\in\mathcal{M}$

步骤 2： 针对算法 3.1 确定的 RIS，利用频谱感知技术确定该 RIS 的状态：若空闲，设备采用传输模式一；否则，采用传输模式二

步骤 2: 执行 $\nu_1 z^{\delta}$ 次步骤 3 至步骤 6

步骤 3: 根据 Beta 分布获取集合 $\mathcal{M}$ 中所有 SF 的经验成功概率：$\hat{\theta}_{n,m}\sim\text{Beta}\left(\alpha_{n,m}(t)+1,\beta_{n,m}(t)+1\right)$

步骤 4: 确定当前的 SF：$I''_{n,t}=\arg\max_{m\in\mathcal{M}} c_m\times\hat{\theta}_{n,m}$

步骤 5： 利用 SF $I''_{n,t}$ 进行传输，并观测传输反馈 $X_{I''_{n,t}}(t)$

步骤 6： 更新动作 $I''_{n,t}$ 上的 Beta 分布参数（后验更新）：当 $X_{I''_{n,t}}(t)=1$ 时，$\alpha_{n,I''_{n,t}}(t)=\alpha_{n,I''_{n,t}}(t-1)+1$；否则，$\beta_{n,I''_{n,t}}(t)=\beta_{n,I''_{n,t}}(t-1)+1$

步骤 7： 确定当前最好的 SF：$c_n^*=\arg\max_{m\in\mathcal{M}} c_m\alpha_{n,m}/\left(\alpha_{n,m}+\beta_{n,m}\right)$

最后，基于最优 RIS 分配策略和最优 SF 分配策略，提出一种 E2Boost 算法来求解该两阶段的 MPMAB 问题，如算法 3.3 所示。从中可以看到，E2Boost 算法分时间段执行（即 $z=1,2,\ldots,Z$）。在每个时间段内，算法序贯地执行最佳 RIS 分配和 SF 分配，从而获得当前时间段内最佳的 RIS 和 SF。值得注意的是，在算法 3.1 的步骤 3 中，传输速率是从集合 $\mathcal{M}$ 中随机选择的。在算法 3.3 的步骤 4 中，当前最佳的 SF 可以反馈至算法 3.1 的步骤 3 中作为当前最佳的传输速率，从而提高传输

效率。最后，E2Boost 算法在每一个 IoT 设备上单独执行，因此是一个完全分布式的算法。

算法 3.3 基于 E2Boosts 算法联合分配策略（在 IoT 设备 n 上执行）

步骤 1：	初始化算法 1-1 和算法 1-2 中的参数
步骤 2:	执行 $z=1,2,\ldots,Z$ 次以下步骤
步骤 3:	最佳 RIS 分配策略，即**算法 3.1**
步骤 4:	最佳 SF 分配策略，即**算法 3.2**

3.4 理论分析

下面推导 E2Boost 算法的伪遗憾的期望上界。根据 3.2.4 节中关于 MPMAB 问题模型描述，令 $\mathbf{a}$ 表示网络中所有 IoT 设备关于 RIS 和 SF 的联合选择的策略，则两阶段 MPMAB 试图解决以下问题

$$\mathbf{a}^* = \arg\max_{\mathbf{a}} \sum_{n=1}^{N} \hat{\mu}_{n,a_n} = \arg\max_{\mathbf{a}} \sum_{n=1}^{N} \mathbb{E}\left[c_{\mathbf{a}} \eta_k(\mathbf{a}) X_{\mathbf{a}}(t)\right] \tag{3-25}$$

其中，$\mathbf{a}^* = \{a_1^*, a_2^*, \cdots, a_N^*\}$ 表示最优的动作策略。根据文献[67]，算法的累积遗憾可以表示成

$$\mathcal{R}eg \triangleq \sum_{t=1}^{T}\sum_{n=1}^{N} r_{n,a_n^*}(t) - \sum_{t=1}^{T}\sum_{n=1}^{N} r_{n,a_n}(t) \tag{3-26}$$

其中，$a_n^* \in \mathbf{a}^*$ 和 T 表示总的时隙数。为了分析方便，下面给出伪遗憾的定义

$$\begin{aligned} \overline{\mathcal{R}eg} &= \sum_{n=1}^{N}\left(T \times \mu_{n,a_n^*} - \mathbb{E}\sum_{t=1}^{T} \mu_{n,a_n}\right) \\ &= \begin{cases} \sum_{n=1}^{N}\sum_{i=1}^{K\times M} \Delta_{n,i}\mathbb{E}[W_{n,i}], & \text{模式一,} \\ \sum_{n=1}^{N}\sum_{i=1}^{M} \Delta_{n,i}\mathbb{E}[W_{n,i}], & \text{模式二,} \end{cases} \end{aligned} \tag{3-27}$$

其中，$\Delta_{n,i} = \mu_{n,a_n^*} - \mu_{n,i}$ 以及 $W_{n,i}$ 表示动作直至时隙 T 被选中的次数。此外，符号 $\mu_{n,i}$ 表示 IoT 设备 n 选择动作 i 得到的平均吞吐量。

可以看到，由于 IoT 设备在该异构网络中有两种传输模式，伪遗憾也由两部分

构成。对于模式一，E2Boost 算法拥有与 GoT（Game of Throne）算法相似的结构。因此，其性能分析可以参照文献[67]。但与 GoT 算法相比，E2Boost 算法具有以下特点：第一，E2Boost 算法是两阶段的 MPMAB 框架，其需要探索的动作空间远远小于 GoT 算法；第二，E2Boost 算法在估计各 RIS 的传输成功概率时，采用了自适应的ε-贪婪算法，来有效权衡 EE 困境，加快收敛速度；最后，E2Boost 算法在进行 SF 分配时，采用了 TS 算法，从而能更快地收敛到最佳的 SF。对于模式二，IoT 设备直接传输至 BS，不存在分布式 RIS 分配的问题，看作传统的随机 MAB 问题，因此可以借鉴文献[84]来分析 TS 算法的伪遗憾上界。

综述所述，E2Boost 算法的伪遗憾上界可由定理 3.1 给出。

定理 3.1. 令$\Gamma_{\max}=\max_{n,i}\mu_{n,i}$表示所有玩家中最大的期望奖励。对于任意的异构上行网络，在给定$\nu_1>0,\nu_2>0,\nu_3>0,\delta\geq 0,0<\varpi<1$和足够小的$\varepsilon$的情况下，E2Boost 算法的伪遗憾上界可以表示为

$$\begin{aligned}\overline{Reg}\leq & N\Gamma_{\max}(1-P_a)\left(2(\nu_1+\nu_2)\log_2^{1+\delta}\left(\frac{T}{\nu_3}+2\right)+(6NK+1)\nu_3\log_2\left(\frac{T}{\nu_3}+2\right)\right)\\ &+P_a(1+\varpi)\sum_{n=1}^{N}\sum_{a_n\in\mathcal{M}}\frac{\log_2 T}{\mathrm{D}_{\mathrm{KL}}(a_n,a_n^*)}\Delta_{n,a_n}\end{aligned} \tag{3-28}$$

其中，$\mathrm{D}_{\mathrm{KL}}(\cdot)$表示 Kukkback-Leibler 散度；符号P_a表示蜂窝网用户活跃的概率。

证明：首先，考虑传输模式一的情况。令E_z表示最优的 RIS 分配策略不在第z个时间段内的事件，其概率可以表示为

$$\begin{aligned}\Pr(E_z)&=\Pr\left(E_z^{k^*},E_z^{m^*}\right)+\Pr\left(E_z^{k^*},\overline{E_z^{m^*}}\right)+\Pr\left(\overline{E_z^{k^*}},E_z^{m^*}\right)\\ &=\Pr\left(E_z^{k^*}\right)+\Pr\left(\overline{E_z^{k^*}},E_z^{m^*}\right)\end{aligned} \tag{3-29}$$

其中，$E_z^{k^*}$表示最优的 RIS 不在算法 3.1 中第z个时间段内的事件；$E_z^{m^*}$表示最优的 SF 不在算法 3.2 中第z个时间段内的事件。

首先，考虑事件$E_z^{k^*}$的概率，即

$$\Pr\left(E_z^{k^*}\right)\leq\Pr\left(\bigcup_{j=0}^{\lfloor z/2\rfloor}P_{e,z-j}\right)+P_{g,z} \tag{3-30}$$

其中，$P_{e,z}$表示在算法 3.1 的 EE 阶段得到的分配结果不是最优分配$\mathbf{a}^*$的概率；$P_{g,z}$表示算法 3.1 的博弈阶段的分配结果不是最优分配$\mathbf{a}^*$的概率。假设$X_{n,k}$服从独立同分布，且每个 IoT 设备在初始阶段均匀地探索所有 RIS。根据文献[67]，$P_{e,z}$可以表示为

$$P_{e,z} \le 2NKe^{-w\nu_1\left(\frac{z}{2}\right)^{\delta} z} + NKe^{-\frac{\nu_1\left(\frac{z}{2}\right)^{\delta}}{36K^2} z} \tag{3-31}$$

其中，w是一个正整数。因此，对多个时间段z累加，可以得到

$$\Pr\left(\bigcup_{j=0}^{\left\lfloor \frac{z}{2} \right\rfloor} P_{e,z-j}\right) \le \frac{2NKe^{-\frac{w}{2}\nu_1\left(\frac{z}{4}\right)^{\delta} z}}{1-e^{-w\nu_1\left(\frac{z}{4}\right)^{\delta}}} + \frac{NKe^{-\frac{\nu_1\left(\frac{z}{4}\right)^{\delta}}{72K^2} z}}{1-e^{-\frac{\nu_1}{36K^2}\left(\frac{z}{2}\right)^{\delta}}} \tag{3-32}$$

接着，令$v^{z*}=[\mathbf{a}^{k*}, C^N]$表示第$z$个博弈阶段的最佳策略，以及$F_z(v^*)$表示在最近$\lfloor z/2 \rfloor+1$阶段内最佳策略被访问的次数，即

$$F_z(v^*) \triangleq \sum_{i=z-\left\lfloor \frac{z}{2} \right\rfloor}^{z} \sum_{t\in\mathcal{G}_z} \mathbb{I}\left(v(t)=v^{i*}\right), \forall k \in \mathcal{K} \tag{3-33}$$

同时，令$\pi_{v^*} = \min\limits_{z-\lfloor z/2 \rfloor \le j \le z} \pi_{v^{i*}}$表示最优策略的平稳分布。在给定$0<\eta<1/2$，且$\pi_{v^*} > 1/\left(2(1-\eta)\right)$情况下，$P_{g,z}$可以表示为

$$P_{g,z} \triangleq \Pr\left(F_z(v^*) \le \frac{1}{2}\sum_{i=z-\left\lfloor \frac{z}{2} \right\rfloor}^{z} \nu_2 i^{\delta}\right) \le \left(C_0 e^{-\frac{\nu_2\eta^2}{144T_m\left(\frac{1}{8}\right)}\left(\pi_{v^*}-\frac{1}{2(1-\eta)}\right)\left(\frac{z}{2}\right)^{\delta}}\right)^z \tag{3-34}$$

其中，C_0是一个依赖于z，π_{v^*}和η的常数。

其次，考虑事件$\left(\overline{E_z^{k^*}}, E_z^{m^*}\right)$的概率。令$P_{t,z}^n$表示玩家$n$没有选择最优 SF 的概率，则条件概率$\Pr\left(E_z^{m^*} \mid \overline{E_z^{k^*}}\right)$可以表示为

$$\begin{aligned}\Pr\left(E_z^{m^*}\mid\overline{E_z^{k^*}}\right)=\Pr\left(\bigcup_{n=1}^{N}P_{t,z}^{n}\right)&\triangleq\sum_{n=1}^{N}\Pr\left(\sum_{m=1,m\neq m^*}^{M}\sum_{j=1}^{z}W_{n,m}^{j}\geq\frac{1}{2}\sum_{i=1}^{z}\nu_3 2^i\right)\\&\overset{(a)}{\leq}\sum_{n=1}^{N}2^{-D_{\mathrm{kl}}\left(\left(1-\frac{c_m^*\theta_{n,m}^*}{\sum_{m=1}^{M}c_m\theta_{n,m}}\right)\middle\|\frac{1}{2}\right)\sum_{i=1}^{z}\nu_3 2^i}\\&\overset{(b)}{\leq}\sum_{n=1}^{N}2^{-2\left(\frac{c_m^*\theta_{n,m}^*}{\sum_{m=1}^{M}c_m\theta_{n,m}}-\frac{1}{2}\right)^2\left(2^{z+1}-2\right)\nu_3}\\&\overset{(c)}{\leq}N2^{-\frac{(M-2)^2\left(2^z-1\right)\nu_3}{M^2}}\end{aligned}\tag{3-35}$$

其中，W_{n,m^*}^{j}表示直至时间段 j 玩家 n 选择次优动作 SF m 的次数；不等式（a）成立是利用了大数定理[113]；不等式（b）成立是利用了 Pinsker 不等式，即 $D_{\mathrm{kl}}(p\|q)\geq 2(p-q)^2$；而不等式（c）成立是因为 $Mc_m^*\theta_{n,m}^*\geq\sum_{m=1}^{M}c_m\theta_{n,m}$。根据关系式 $\Pr\left(\overline{E_z^{k^*}},E_z^{m^*}\right)=\Pr\left(E_z^{m^*}\mid\overline{E_z^{k^*}}\right)\Pr\left(\overline{E_z^{k^*}}\right)$，可以得到

$$\begin{aligned}\Pr\left(\overline{E_z^{k^*}},E_z^{m^*}\right)\leq N2^{-\frac{(M-2)^2\left(2^z-1\right)\nu_3}{M^2}}&\left(1-\frac{2NKe^{-\frac{w}{2}\nu_1\left(\frac{z}{4}\right)^{\delta}z}}{1-e^{-w\nu_1\left(\frac{z}{4}\right)^{\delta}}}+\frac{NKe^{-\frac{\nu_1\left(\frac{z}{4}\right)^{\delta}}{72K^2}z}}{1-e^{-\frac{\nu_1}{36K^2}\left(\frac{z}{2}\right)^{\delta}}}\right.\\&\left.+\left(C_0e^{-\frac{\nu_2\eta^2}{144T_m(\frac{1}{8})}\left(\pi_{\nu^*}-\frac{1}{2(1-\eta)}\right)\left(\frac{z}{2}\right)^{\delta}}\right)^z\right)\end{aligned}\tag{3-36}$$

接着，结合表达式式（3-31）、（3-34）和（3-36），可以得到关于算法在时间段 z 的伪遗憾 $\overline{Reg}_z$ 为

$$
\begin{aligned}
\overline{Reg}_z &\le N\Gamma_{\max}\nu_2 z^{\delta} + \Pr(E_z)N\Gamma_{\max}\nu_1 z^{\delta} + \Pr(E_z)N\Gamma_{\max}\nu_3 2^z \\
&\le N\Gamma_{\max}\nu_2 z^{\delta} + N\Gamma_{\max}\left(\nu_1 z^{\delta} + \nu_3 2^z\right)\left[\frac{2NKe^{-\frac{w}{2}\nu_1\left(\frac{z}{4}\right)^{\delta} z}}{1-e^{-w\nu_1\left(\frac{z}{4}\right)^{\delta}}}\right. \\
&\quad \left. + \frac{NKe^{-\frac{\nu_1\left(\frac{z}{4}\right)^{\delta}}{72K^2}z}}{1-e^{-\frac{\nu_1}{36K^2}\left(\frac{z}{2}\right)^{\delta}}} + \left(C_0 e^{-\frac{\nu_2\eta^2}{144T_m\left(\frac{1}{8}\right)}\left(\pi_{v^*}-\frac{1}{2(1-\eta)}\right)\left(\frac{z}{2}\right)^{\delta}}\right)^z\right] \\
&\quad + N^2\Gamma_{\max}\left(\nu_1 z^{\delta} + \nu_3 2^z\right)2^{-\frac{(M-2)^2\left(2^z-1\right)\nu_3}{M^2}}\left[1-\frac{2NKe^{-\frac{w}{2}\nu_1\left(\frac{z}{4}\right)^{\delta} z}}{1-e^{-w\nu_1\left(\frac{z}{4}\right)^{\delta}}}\right. \\
&\quad \left. + \frac{NKe^{-\frac{\nu_1\left(\frac{z}{4}\right)^{\delta}}{72K^2}z}}{1-e^{-\frac{\nu_1}{36K^2}\left(\frac{z}{2}\right)^{\delta}}} + \left(C_0 e^{-\frac{\nu_2\eta^2}{144T_m\left(\frac{1}{8}\right)}\left(\pi_{v^*}-\frac{1}{2(1-\eta)}\right)\left(\frac{z}{2}\right)^{\delta}}\right)^z\right] \\
&\le N\Gamma_{\max}\left[\left(\frac{\nu_1}{2}+3NK+\nu_2\right)z^{\delta} + (6NK+1)\nu_3\right] + N^2\Gamma_{\max}\left(\nu_1 z^{\delta} + \nu_3 2^z\right)2^{-\nu_3\left(2^z-1\right)}
\end{aligned}
\tag{3-37}
$$

其中，第一个不等式成立是考虑了最坏情况，即每个玩家都产生最大的遗憾 $\Gamma_{\max}$；第二个不等式成立是利用（3-31）和（3-34）；最后一个不等式成立是因为

$$
\max\left\{C_0 e^{-\frac{\nu_2\eta^2}{144T_m\left(\frac{1}{8}\right)}\left(\pi_{v^*}-\frac{1}{2(1-\eta)}\right)\left(\frac{z}{2}\right)^{\delta}}, e^{-\frac{w}{2}\nu_1\left(\frac{z}{4}\right)^{\delta}}, e^{-\frac{\nu_1\left(\frac{z}{4}\right)^{\delta}}{72K^2}}\right\} < \frac{1}{2} \tag{3-38}
$$

以及

$$
2^{-\frac{(M-2)^2\left(2^z-1\right)\nu_3}{M^2}} \le 2^{-\nu_3\left(2^z-1\right)} \tag{3-39}
$$

最后，通过对时间段 z 进行累加，算法在传输模式一下的总伪遗憾可以表示为

$$\begin{aligned}\overline{Reg}^{(1)} &\overset{(a)}{\le} \sum_{z=1}^{Z}\overline{Reg}_z \overset{(b)}{\le} N\Gamma_{\max}\sum_{z=1}^{z_0}\left((\nu_1+\nu_2)z^{\delta}+\nu_3 2^z\right)\\ &+N\Gamma_{\max}\sum_{z=z_0+1}^{Z}\left(N\Gamma_{\max}\left(\frac{\nu_1}{2}+3NK+\nu_2\right)z^{\delta}n\right.\\ &\left.+N\Gamma_{\max}(6NK+1)\nu_3+N^2\Gamma_{\max}\left(\nu_1 z^{\delta}+\nu_3 2^z\right)2^{-\nu_3\left(2^z-1\right)}\right)\\ &\overset{(c)}{\le} N\Gamma_{\max}\sum_{z=1}^{z_0}\nu_3 2^z+N\Gamma_{\max}\sum_{z=1}^{Z}(\nu_1+\nu_2)z^{\delta}+ZN\Gamma_{\max}(6NK+1)\nu_3\\ &\overset{(d)}{\le} N\Gamma_{\max}(\nu_1+\nu_2)\log_2^{1+\delta}\left(\frac{T}{\nu_3}+2\right)+N\Gamma_{\max}\nu_3 2^{z_0+1}\\ &+N\Gamma_{\max}(6NK+1)\nu_3\log_2\left(\frac{T}{\nu_3}+2\right)\end{aligned} \tag{3-40}$$

其中，不等式（b）成立是利用了（3-37）；不等式（d）成立是利用了 $\sum_{z=1}^{Z}z^{\delta}\le Z^{1+\delta}$，$T\ge\sum_{z=1}^{Z}\nu_3 2^z\ge\nu_3(2^Z-2)$ 以及 $Z^{1+\delta}\le\log_2^{1+\delta}\left(T/\nu_3+2\right)$。

在传输模式二下，算法的伪遗憾可以由文献[84]给出，即

$$\overline{Reg}^{(2)}\le P_a(1+\varpi)\sum_{n=1}^{N}\sum_{a_n\in\mathcal{M}}\frac{\log_2 T}{\mathrm{D}_{\mathrm{KL}}(a_n,a_n^*)}\Delta_{n,a_n} \tag{3-41}$$

其中，$\varpi\in(0,\,1)$ 以及 P_a 是蜂窝网用户的活跃概率。综上所述，E2Boost 算法的伪遗憾上界可以表示为

$$\begin{aligned}\overline{Reg} &= \overline{Reg}^{(1)}+\overline{Reg}^{(2)}\\ &\le N\Gamma_{\max}(1-P_a)\left(2(\nu_1+\nu_2)\log_2^{1+\delta}\left(\frac{T}{\nu_3}+2\right)+(6NK+1)\nu_3\log_2\left(\frac{T}{\nu_3}+2\right)\right)\\ &+P_a(1+\varpi)\sum_{n=1}^{N}\sum_{a_n\in\mathcal{M}}\frac{\log_2 T}{\mathrm{D}_{\mathrm{KL}}(a_n,a_n^*)}\Delta_{n,a_n}.\end{aligned} \tag{3-42}$$

因此，定理 3.1 证毕。 □

注 1. 定理 3.1 的前两项是传输模式一下的伪遗憾上界；而第三项是传输模式二的伪遗憾上界。而且，这两部分的权重依赖于蜂窝网用户的活跃概率 P_a。

注 2. 定理 3.1 表明 E2Boost 算法的累积遗憾呈对数增长。当时间 T 足够大时，每个时隙所产的遗憾将趋于 0。换言之，E2Boost 算法能收敛到最佳的 RIS 和 SF。

注 3. 定理 3.1 还表明 E2Boost 算法的伪遗憾上界远远小于 GoT 算法。根据文献

[67]，GoT 算法的伪遗憾上界为

$$\begin{aligned}\overline{Reg}_{\mathrm{GoT}} \le & 4N\Gamma_{\max}(\nu_1+\nu_2)\log_2^{1+\delta}\left(\frac{T}{\nu_3}+2\right) \\ & + N\Gamma_{\max}(6NKM+1)\nu_3\log_2\left(\frac{T}{\nu_3}+2\right)=\mathcal{O}(\log_2^{1+\delta}T)\end{aligned} \tag{3-43}$$

假设 $\nu_1=\nu_2=\nu_3$，且 $\delta=0$，可以得到 E2Boost 算法的伪遗憾上界为

$$\overline{Reg} \le (5+6NK)\nu_1 N\Gamma_{\max}\log_2\left(\frac{T}{\nu_3}+2\right) \tag{3-44}$$

以及 GoT 算法的伪遗憾上界为

$$\overline{Reg}_{\mathrm{GoT}} \le (9+6NKM)\nu_1 N\Gamma_{\max}\log_2\left(\frac{T}{\nu_3}+2\right) \tag{3-45}$$

可以看出，$\overline{Reg}$ 比 $\overline{Reg}_{\mathrm{GoT}}$ 大概低了 M 倍。这个结果可以利用下一节的仿真结果得到验证。

3.5 仿真结果

本节通过数值仿真来验证所提算法的有效性。仿真参数的选择主要参考文献[114]，[115]和 3GPP 标准[109]，且所有的结果均由1000次蒙特卡洛仿真得到。

3.5.1 参数配置与对比算法

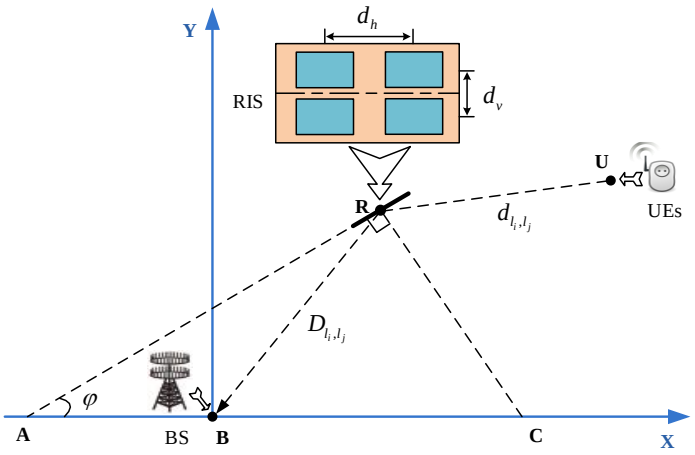


图 3.2 RIS 的放置示意图（俯视）

表 3.1 仿真参数设置

符号	物理意义	数值
Ω_n	IoT 设备 n 的发送功率	20 dBm
$\sigma_\omega^2+\sigma_y^2$	背景噪声功率+信号干扰功率	-95 dBm
f_c	载波中心频率	5.9 GHz
B	信号带宽	40 MHz
G	天线增益	1
ζ	莱斯因子	4
P_a^k	蜂窝网用户在 RIS k 的活跃概率	0.2
A	RIS 的反射系数	1
b	每个 RIS 元素上 PIN 二极管数目	8

图 3.2 是 RIS 的放置示意图。一个 RIS 上有101×101个元素，每个元素与 x 轴方向成的角度为$\angle\varphi$，则各元素的三维坐标可以用以下公式得到

$$\begin{cases} x(l_1,l_2)=(l_1-51)d_v\cos\angle\varphi+x_{\mathrm{R}}, \\ y(l_1,l_2)=(l_1-51)d_v\sin\angle\varphi+y_{\mathrm{R}}, \\ z(l_1,l_2)=(l_2-51)d_h+10, \end{cases} \tag{3-46}$$

其中，$d_v=d_h=0.01$分别表示 RIS 在水平方向和垂直方向每个元素的间距。此外，在仿真中，考虑了两种不同的相移设置：最优相移和随机相移。其中，最优相移是针对网络中的蜂窝网用户设置的，可以由文献[114]给出

$$\tau_{l_1,l_2}=\left\lfloor\left(\Pi-\frac{2\pi}{\lambda}L^k_{l_1,l_2}\right)\frac{2^b}{2\pi}\right\rfloor\frac{2\pi}{2^b} \tag{3-47}$$

其中，Π 是任意给定的一个常数；$L^k_{l_1,l_2}$ 表示蜂窝网用户到基站并经过 RIS k 上第 (l_1,l_2) 个元素的距离；对于随机相移，假设 RIS 上所有元素的相移都相等，且随机地从$[0,2^b-1]$之间选择一个。在随后的仿真中，设置随机相移为170。另外，关于扩频因子的参数由表 3.1 给出。在实际系统中，每台 IoT 设备有 6 个扩频因子，分别为$(7,8,9,10,11,12)$。在仿真中，设定 6 个扩频因子对应的判决门限，通过与接收到的 SINR 进行比较，来得到设备该次传播是否成功的反馈，最后通过求平均便可以获得设备在某一 SF 上的传输成功概率。

表 3.2 扩频因子的相关参数设置

扩频因子（SF）	7	8	9	10	11	12
传输速率（Mbps）	1.09	0.63	0.35	0.20	0.11	0.06
判决门限（$\times 10^3$）	4.5	4	3.5	3	2.5	2

本节比较 E2Boost 算法与几种不同算法的性能，比如最优解、GoT 算法、Q-Learning 算法、随机选择算法、无 WD 距离的 E2Boost 算法和无 TS 的 E2Boost 算法。几种算法的设置具体如下：

- 最优解：其通过求解两阶段的 MPMAB 问题得到。在知道每个 RIS 的估计的成功概率的情况下（上帝视角），可以通过匈牙利算法为每个 IoT 设备分配最优的 RIS。在仿真中，估计的传输成功概率通过10^5次蒙特卡洛的信道仿真得到。具体地，利用表 3.2 中的传输速率和其对应的解调门限，可以得到相应的传输反馈（成功或失败），然后计算得到成功概率。
- GoT 算法：它是一种完全分布式的算法，且与所提算法拥有相同框架。不同的是 GoT 算法中没有ε-贪婪算法和 TS 算法来进一步权衡 EE 困境。此外，GoT 算法需要探索的动作空间远大于 E2Boost 算法，因为后者分开探索 RIS 和 SF，而前者探索 RIS 与 SF 的组合。因此，只要在仿真中注意两种算法不同点，便可实现 GoT 算法。
- Q-Learning 方法：在该方法中，状态是 RIS 是否被蜂窝网用户占用，动作是 RIS 或者 SF，状态转移概率为蜂窝网用户的活跃概率。
- 随机选择方法：在传输模式一下，每个 IoT 设备随机地从集合$\mathcal{K}\otimes\mathcal{M}$中选择一对 RIS 和 SF；而在传输模式二下，每个 IoT 设备随机地从集合$\mathcal{M}$中选择一个 SF。
- 无 WD 距离的 E2Boost 算法：在 E2Boost 算法的基础，去除自适应调整 EE 参数ε的步骤。
- 无 TS 的 E2Boost 算法：在 E2Boost 算法的基础上，在最优 SF 分配中去除 TS 算法。

3.5.2 静态网络场景仿真

图 3.3 给出的是一个静态网络的场景示意图，其中，3 个 IoT 设备位于半径为 45 m 的圆形区域内。假设所有的蜂窝网用户均位于圆形区域的中心，即坐标为 $(x,y)=(150,150)$ m。在圆形区域外是 3 个 RIS 和位于原点的 BS。根据各个设备和用户的坐标，利用欧氏距离公式可以得到相应的距离，如 D_{l_i,l_j} 和 d_{l_i,l_j}。此外，假设 RIS 和 BS 的高度分布为10 m 和 20m。

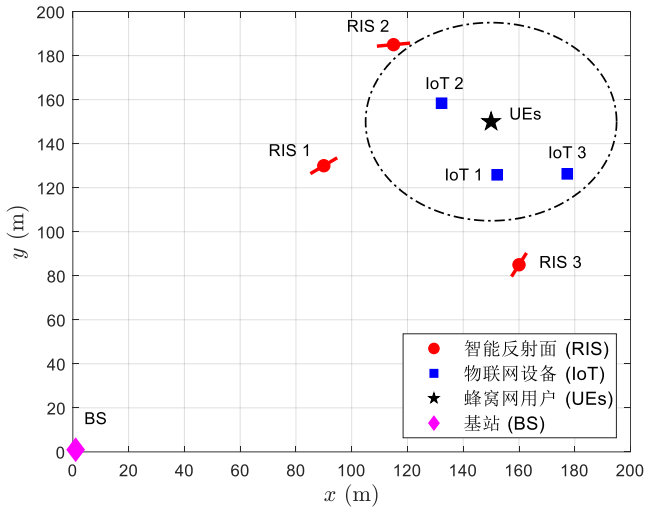


图 3.3 一个静态网络场景图（俯视）

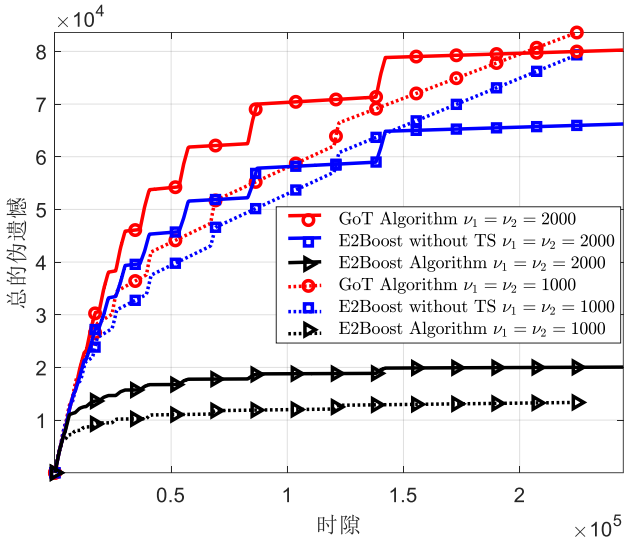


图 3.4 在最佳相移设定下 E2Boost 算法、没有 TS 的 E2Boost 算法和 GoT 算法的伪遗憾随时隙变化的曲线

图 3.4 描述了 E2Boost 算法、没有 TS 的 E2Boost 算法和 GoT 算法在最佳相移设定下的伪遗憾随着时隙变化的曲线。其中，算法的参数设定为：$\nu_1=\nu_2=1000$，$\nu_1=\nu_2=2000$，$\nu=1.4$，$\delta=0$，$\varepsilon=0.01$，$Z=10$。从图中可以看到，所提出的算法在这两种情况下都具有最低的伪遗憾，其原因是 E2Boost 算法需要探索的动作空间远小于其它算法（因为采用了两阶段分配机制）。此外，E2Boost 算法的伪遗憾在$\nu_1=\nu_2=1000$情况下远低于$\nu_1=\nu_2=2000$情况。一方面，ν_1和ν_2的值越大，算法需要探索所有动作的时间就越长，从而导致性能损失。另一方面，当$\nu_1=\nu_2=1000$时，GoT 算法和没有 TS 的 E2Boost 算法因为ν_2的值太小而无法在博弈阶段解决 IoT 设备之间的冲突，从而导致算法无法收敛。从图中还可以看到，GoT 算法的伪遗憾大约是 E2Boost 算法的 4 倍，这验证前面的理论分析结果。

图 3.5 比较了 E2Boost 算法、$\epsilon=0$和$\epsilon=1$（没有 WD）的 E2Boost 算法、没有 TS 的 E2Boost 算法、GoT 算法和随机选择方法在最佳相移设置下的性能比较曲线，其中，$\nu_1=\nu_2=2000,\ Z=10$。从图中可以看到，所提算法优于其他算法，且接近最优解。相比之下，因为在第一阶段没有任何探索，$\epsilon=0$的 E2Boost 算法的性能最差。同时，$\epsilon=1$的 E2Boost 算法和没有 TS 的 E2Boost 算法优于 GoT 算法，这表明自适应的ε-贪婪算法可以有效提高所提算法的性能。最重要的是，由于采用了两阶段分配机制，E2Boost 算法比 GoT 算法和没有 TS 的 E2Boost 算法具有更快的收敛速度。

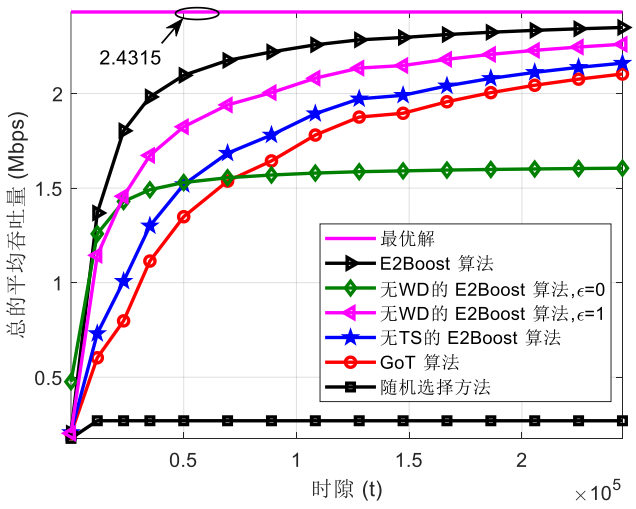


图 3.5 在最佳相移设定下所提算法与比较算法的网络平均吞吐量随时隙变化曲线

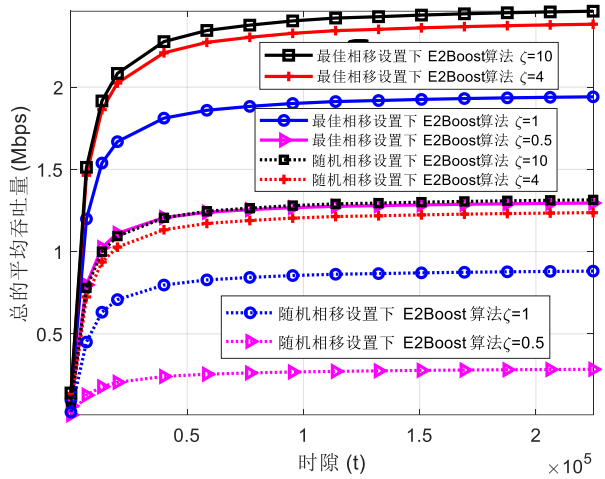


图 3.6 在不同莱斯因子和相移设定下所提算法与比较算法的性能曲线

考虑 RIS 信道模型对所提算法的影响，图 3.6 描绘了 E2Boost 算法在不同 Rice 因子 $\zeta = 0.5,\ 1,\ 4,\ 10$ 与相移设定下的性能曲线，其中，$\nu_1 = \nu_2 = 2000,\ Z = 10$。从图中可以看到，对于不同的 Rice 因子，E2Boost 算法在最佳相移设置的性能比随机相移设置要好得多。这是因为最优相移是为蜂窝网用户设置，因此，靠近用户的 IoT 设备也将具有更好的性能。另一方面，更大的 ζ 将导致更高的平均总吞吐量。这种现象可以用莱斯信道模型解释，其中较大的 ζ 意味着信道增益由 LoS 分量控制，即 IoT-RIS-BS 路径。因此，当 ζ 趋向 $+\infty$ 时，信道增益由 RIS 主导；而 ζ 趋向于 0 意味着 IoT 设备仅在模式二上传输。

3.5.3 动态网络场景仿真

下面评估动态网络场景中所提算法的性能。首先考虑一个简单的动态网络场景，即圆形内的 3 个 IoT 设备的位置在每次蒙特卡洛仿真时随机变化，具体参数和要求与图 3.2 中相同。另外，两个 IoT 设备之间的距离不少于 5 m，且 RIS，BS 和蜂窝网用户的位置参数与图 3.2 中一致。

图 3.7 比较了在上述网络场景与最佳相移设置下不同算法的平均总吞吐量随时隙变化的曲线，其中 $\nu_1 = \nu_2 = 2000$，$Z = 11$。从图中可以看到，除了随机选择方法外，其它所有算法的性能都随着时隙增加而增加。同样，E2Boost 算法在其中具有最好的性能和最快的收敛速度。此外，Q-Learning 方法也表现出较快的收敛速度，但由于缺乏非合作博弈阶段来解决玩家之间的分配冲突，导致其性能下降。与静态

网络场景下的图 3.5 相比，最优解与这些算法之间的差距变大了。这是因为在固定参数 ν_2 和时间范围 T 时，它们无法从某些极端的网络场景中实现最佳的 RIS 分配。

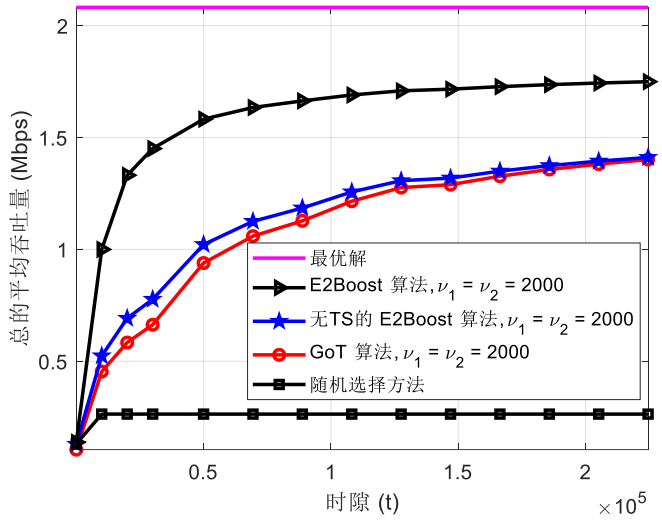


图 3.7 在最佳相移设定和动态网络场景下所提算法与比较算法的性能曲线

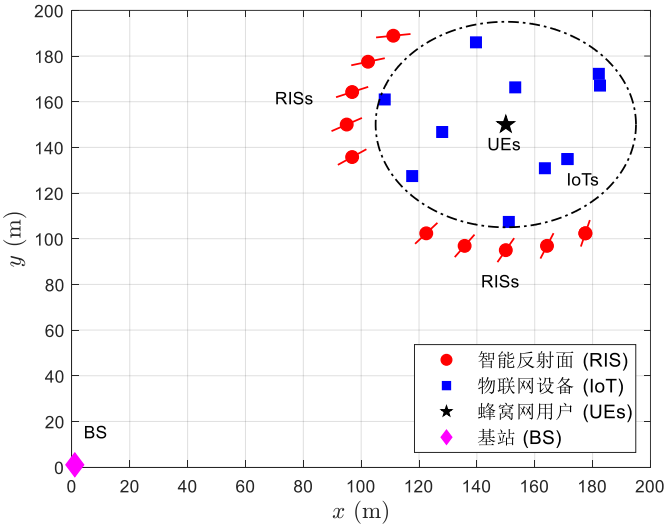


图 3.8 一个动态网络场景图（俯视）

最后，考虑网络中 IoT 设备数目对所提算法的影响。在该网络场景中，RIS 的数目为10，且等间隔分布在半径为55m 的圆形区域的 $3\pi/4$ 到 $5\pi/4$ 的弧形区域，如图 3.8 所示。图 3.9 给出了最优相移设置下最优解、E2Boost 算法、GoT 算法和随机选择方法的性能曲线，其中，$\nu_1=\nu_2=2000, K=10, Z=10$。从图中可以看到，

所有算法的性能随着 IoT 设备数量增加而提高。因为两阶段分配机制，所提算法每次只需探索较小的动作空间，所以 E2Boost 算法的性能优于 GoT 算法和随机选择方法。此外，最优解与这些算法之间的性能差距随着 IoT 设备的数量增加而增加。这是因为随着 IoT 设备数量的增加，设备之间的碰撞概率也随之增加，从而导致了很多的性能损失。

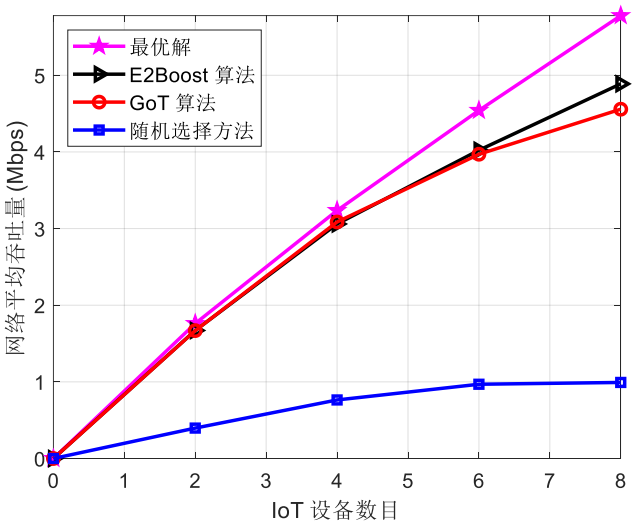


图 3.9 在最佳相移设定和网络场景图 3.8 下所提算法与比较算法的性能随着 IoT 数目变化的曲线

3.6 本章小结

本章研究了分布式异构网络中的资源分配问题，其系统目标是通过为每个 IoT 设备分配最佳的 RIS 和 SF 来最大化网络中所有 IoT 设备的传输速率之和。由于 IoT 设备无法准确获取 RIS 的信息，以及无中心节点协调设备之间的信息交互，该速率最大化问题被建模为两阶段 MPMAB 框架，其动作在第一阶段和第二阶段分别是 RIS 和 SF。然后，结合非合作博弈理论和 MAB 技术，提出了一种 E2Boost 算法来求解该两阶段 MPMAB 问题。一方面，由于采用自适应 ε-贪婪算法和 TS 算法，E2Boost 算法可以有效地权衡 EE 困境；另一方面，由于采用了非合作博弈理论，每个 RIS 可以保证收敛到最佳的 RIS。此外，推导了 E2Bosst 算法伪遗憾的上界，即 $\mathcal{O}(\log_2^{1+\delta} T)$，其表明当 T 足够大时，算法每回合的遗憾将趋向于 0 。最后，

仿真结果证明了所提算法的有效性。更重要的是，由于采用两阶段分配机制，它对动作空间的大小不敏感，该特点在实际应用中有很大优势。

第 4 章 基于马尔科夫 MAB 的高密度物联网设备调度策略

本章介绍基于马尔科夫 MAB 和信息新鲜度的物联网设备调度策略[①]。在已有工作中，基于信息新鲜度的设备调度问题主要考虑理想、非理想信道、以及独立信源的情况，忽略了信源之间的相关性。本章通过考虑信源之间的相关性，构建了一种基于马尔科夫 MAB 框架的设备调度问题。在理想信道和非理想信道情况下，分别提出了基于广义惠特尔和广义部分惠特尔的调度策略，最小化网络平均信息新鲜度。结合优化理论、马尔科夫决策过程和 MAB 技术，理论推导了所提方法的性能下界。最后，通过实验仿真验证了所提方法的有效性，且表明在高密度网络中所提策略的性能显著优于独立信源下的调度策略。

4.1 引言

随着下一代无线通信系统的广泛部署，物联网（Internet-of-Things, IoT）在电子健康、智能家居、驾驶、监控、语义信息分析等日常生活和行业中发挥着越来越重要的作用[116]。IoT 通常由三部分组成，即物联网设备、传输网络和用于信息融合的基站（Base Station, BS）[117]。IoT 设备，也称为信源，通常由人工或随机地部署在不同位置以感知周围环境的物理特征，如温度、湿度、污染水平等；传输网络的作用是将设备采集样本被传送到 BS 进行信息融合；而 BS 的作用是处理收集的样本数据并提取有意义的信息，对网络做出相应决策。在这一过程中，所提取信息的准确度取决于感知样本的新鲜度，并影响 BS 的相关决策。因此，如何有效处理这些时间敏感的感知样本，并保证 BS 决策的准确度是物联网中的一个重要问题[118]。

近几年，信息新鲜度（Age-of-Information, AoI）作为一种新的网络性能指标引起了广泛关注，其定义为从生成最新的数据包至目的节点所经过的时间[119][120]。与时延、吞吐量等传统性能指标相比，AoI 提供了一个新的视角来量化物联网中样本的新鲜度或准确性。基于 AoI 的调度问题在于如何通过调度 IoT 设备的状态更新来最小化网络的平均 AoT。在文献中，已有大量的工作讨论该了问题。但这些工作

[①] 本章内容已发表于 IEEE Transaction on Wireless Communications。

要么单独考虑理想或非理想信道，要么假设信源之间相互独立；很少有工作联合考虑非理想信道和相关信源下基于 AoI 的设备调度问题。

因此，本章考虑理想、非理想信道、以及相关信源的情况下，基于 AoI 的物联网设备调度问题。首先，将其建模为一个马尔科夫决策过程（Markov Decision Process，MDP）。由于信源之间的相关性，使用传统的值函数迭代或策略迭代方法很难求解该 MDP 问题[23]；而且，其近似解仍然存在计算复杂度高和无法提供性能保证的问题[121][122]。因此，进一步将该调度问题建模为一个 CRMAB（Correlated Restless Multi-Armed Bandit）问题[123]。其中，玩家是 BS，动作是 IoT 设备，奖励和状态都是 AoI。其次，通过为每个动作引入一个待定状态变量，将该高维的 CRMAB 问题解耦成多个一维的 MAB 问题。接着，分别求解每一个 MAB 问题的 Bellman 方程得到惠特尔索引（Whitll Index，WI）闭式表达式。然后，基于该 WI 表达式，推导出理想信道下的广义惠特尔索引（Generalized WI，GWI）调度策略和非理想信道下的广义部分惠特尔索引（Generalized Partial WI，GPWI）调度策略。最后，通过求解松弛的拉格朗日问题得到所提策略的理论性能下界。数值仿真结果验证该理论分析的正确性，且表明在高密度网络中所提策略的性能显著优于独立信源下的调度策略。

4.2 系统模型

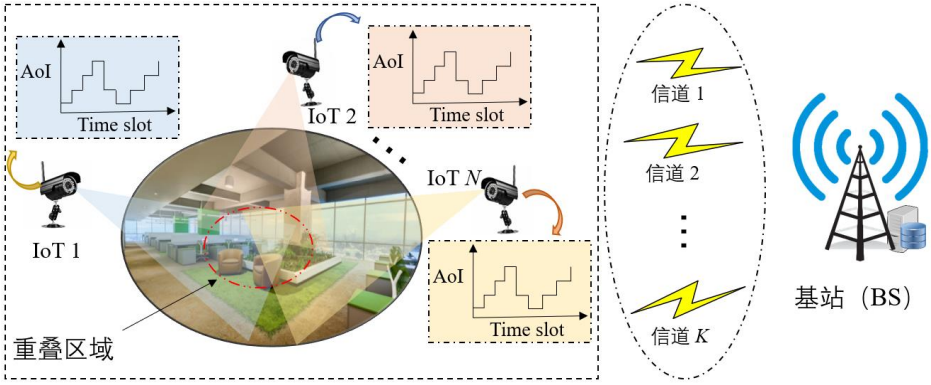


图 4.1 IoT 网络场景示意图

考虑一个 IoT 网络，其中 N 个 IoT 设备部署在一个区域监视物理环境，如图 4.1 所示。这些 IoT 设备需要将包含其状态信息的感知样本通过 K 个信道传输至

BS，然后 BS 根据提取的状态信息做出相应的决策。假设时间是离散化的，即 $t=1,2,\cdots,T$。同时，假设网络中 IoT 设备的数量大于信道数目，且在每个时隙内一个信道最多可以被一个信源（或 IoT 设备）占用。每个时隙开始时，BS 需要决定调度哪些 IoT 设备到 K 个信道中来更新它们的状态。在该网络中，假设信源之间的样本是相关的，即更新某一个信源的状态，其它与其相关的信源也会更新部分信息。系统的目标是每个时隙通过调度 K 个 IoT 设备进行状态更新来最小化时间 T 内网络的平均 AoI。

4.2.1 信道与信源模型

本节考虑两种信道模型：理想信道模型和非理想信道模型。令 p_{nk} 表示信源 n 在信道 k 上的传输成功概率，则信源 n 的信道质量向量可以表示为 $\vec{p}_n=\left(p_{n1},\ldots,p_{nK}\right)$。同时，假设成功概率 p_{nk} 中信源与信道都是相互独立的。在理想信道模型中，信源 k 对于不同的信道有相同的传输成功概率，即 $p_{n1}=\ldots=p_{nK}$。因此，为了方便表示，在该模型下可以省略信道的索引 k。然而，在非理想信道模型下，由于无线信道受到衰落和地理环境的影响，各信道可能经历不同的衰落，所以传输成功概率对于不同的信道和信源均可能不同，即 p_{nk}，$\forall k\in\mathcal{K}$，$\forall n\in\mathcal{N}$ 不同，其中，$\mathcal{K}$ 和 $\mathcal{N}$ 分别表示所有信道和信源的集合。

根据文献[124]，信源之间的相关性可以通过它们的位置和样本来表征，分别对应于空间相关性和时间相关性。下面，通过定义无向图 $\mathcal{G}=\{\mathcal{N},\mathcal{E}\}$ 来刻画物联网中各信源之间的相关性。在图 $\mathcal{G}$ 中，$\mathcal{N}$ 表示顶点集合，每一个顶点代表一个信源；$\mathcal{E}$ 表示边的集合，每一条边代表两个信源之间存在相关性，且可以写成一个 $N\times N$ 矩阵。换言之，$\mathcal{E}=[\mathbf{e}_1;\mathbf{e}_2;\ldots;\mathbf{e}_N]$，其中第 k 行向量中的元素为 $\mathbf{e}_n=[e_{n1},e_{n2},\ldots,e_{nN}]$，且 $\mathcal{E}\in[0,1]^{N\times N}$。另外，元素 e_{ij} 表示第 j 个信源的 AoI 衰减因子，其定义为当 BS 接收到信源 i 的样本时，信源 j 的状态更新程度。因此，矩阵 $\mathcal{E}$ 的对角线元素都等于 1，即 $e_{ii}=1,\forall i\in\mathcal{N}$。值得注意的是，由于 IoT 设备之间的地理位置和硬件精度的差异，矩阵 $\mathcal{E}$ 是不对称的，即 $e_{ij}\neq e_{ji},\forall i,j\in\mathcal{N}$。举一个例子，如图 4.2 所示，其中，重叠区域分别占 IoT 设备 1 和 IoT 设备 2 的监控区域的 50%和 30%。此时，

信源 1 和信源 2 的 AoI 衰减因子分别为$e_{21}=0.5$和$e_{12}=0.3$，也就是说$e_{ij}\neq e_{ji}$。

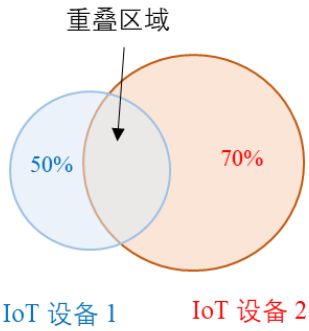


图 4.2 两个 IoT 设备监测区域重叠示意图

4.2.2 信息新鲜度模型

不失一般性，本节采用文献[125]中的即时生成（Generate-at-Will）模型来描述网络中 AoI 的演化过程。在该模型中，无论 IoT 设备何时被调度，它都将立即生成一个数据包进行传输。令$A_n(t)$表示信源n在时隙t的 AoI，则当信源n被调度且传输成功时，它的 AoI 减少到1；而当信源n没有被调度或调度后传输失败，其 AoI 为原 AoI 乘上衰减因子再加上1。综上，信源n随时间变化的 AoI 可以表示为

$$A_n(t+1)=\begin{cases}1, & \text{成功},\\ A_n(t)\alpha_n(t)+1, & \text{其它},\end{cases} \tag{4-1}$$

其中，$\alpha_n(t)=\prod_{j\in\mathcal{I}_g^t}(1-e_{jn})$表示 BS 调度信源集合$\mathcal{I}_g^t$去更新状态时，信源$n$的残余 AoI 程度。$\mathcal{I}_g^t$表示在时隙$t$被调度且传输成功的信源的集合。例如，在图 4.2 中，假设信源 1 和 2 在时隙t的 AoI 均为10，即$A_1(t)=A_2(t)=10$。若在上一个时隙信源 1 被调度且传输成功，则根据式（4-1），在下一个时隙信源 1 和 2 的 AoI 分别为$A_1(t+1)=1$和$A_2(t+1)=10\times(1-0.3)+1=8$。

令$u_{nk}(t)$表示在时隙t的一个二元决策变量。其中，$u_{nk}(t)=1$表示信源n被调度在信道k上进行数据传输；否则，$u_{nk}(t)=0$。由于每个时隙一个信道最多只能被一个信源占用，因此，$\sum_{n=1}^{N}u_{nk}(t)=1,\forall k\in\mathcal{K}$，则该网络平均 AoI 最小化问题可以建模成

$$\begin{aligned}\min_{\pi}\quad & \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\sum_{n=1}^{N}\mathbb{E}\left[A_n^{\pi}(t)\right]\\ \text{s.t.}\quad & \alpha_n(t)=\prod_{j\in\mathcal{I}_g^t}(1-e_{jn}),\forall n\in\mathcal{N},\forall t\\ & \sum_{n=1}^{N}u_{nk}^{\pi}(t)=1,\forall k\in\mathcal{K}\\ & u_{nk}^{\pi}(t)\in\{0,1\},\forall k\in\mathcal{K},\forall t\end{aligned}\tag{4-2}$$

其中，$\mathbb{E}[\cdot]$表示取期望运算，其操作在随机调度策略π上。下面称问题（4-2）为原问题。

4.2.3 MDP 问题模型

通过分析不难发现，上述调度问题可以看成是一个 MDP 问题。令$\mathcal{A}\triangleq\{A_1(t),\ldots,A_N(t)\}$表示系统的状态空间，即$N$个信源在时隙$t$的 AoI；此外，系统的动作空间可以表示为$\mathcal{U}\triangleq\{u_{1k},\ldots,u_{Nk}\}\in\{0,1\}^N$，其中，$u_{nk}=1$表示信源$n$被调度到信道$k$；否则，$u_{nk}=0$。根据（4-1），系统执行动作$\mathbf{u}_t\in\mathcal{U}$且状态由$\mathbf{a}_t=\{A_1(t),\ldots,A_N(t)\}$转移到$\mathbf{a}_{t+1}=\{A_1(t+1),\ldots,A_N(t+1)\}$的概率为

$$\mathbb{P}_{\mathbf{a}_t|\mathbf{a}_{t+1}}(\mathbf{u}_t)=\begin{cases}\displaystyle\prod_{n,k\in\mathcal{I}^t}p_{nk}, & \text{成功}\\ \displaystyle\prod_{n,k\in\mathcal{I}^t}(1-p_{nk}), & \text{失败}\\ \displaystyle\prod_{n,k\in\mathcal{I}_g^t}p_{nk}\prod_{n,k\in\overline{\mathcal{I}}_g^t}(1-p_{nk}), & \text{其它}\end{cases}\tag{4-3}$$

其中，$\mathcal{I}^t$表示信源n被调度在信道k的(n,k)集合，即$\mathcal{I}^t=\{(n,k)\,|\,u_{nk}=1\}$。另外，$\mathcal{I}_g^t$表示信源$n$被调度在信道$k$且传输成功的$(n,k)$集合；而$\overline{\mathcal{I}}_g^t$表示信源$n$被调度在信道$k$但传输失败的$(n,k)$集合。因此，该 MDP 问题可以描述为

$$\begin{aligned}\min_{\pi\in\mathcal{U}} \quad & \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\sum_{n=1}^{N}\mathbb{E}\left[A_n^{\pi}(t)\right] \\ \text{s.t.} \quad & \alpha_n(t)=\prod_{j\in\mathcal{I}_g^t}(1-e_{jn}),\forall n\in\mathcal{N},\forall t \\ & \sum_{n=1}^{N}u_{nk}^{\pi}(t)=1,\forall k\in\mathcal{K} \\ & u_{nk}^{\pi}(t)\in\{0,1\},\forall k\in\mathcal{K},\forall t\end{aligned} \tag{4-4}$$

从中可以看到，上式与原问题的区别在于调度策略空间不同。但是，由于系统状态的数量不可数或连续，采用传统的值函数迭代或策略迭代方法很难解决这个问题。在已有文献中，部分工作通过离散化状态空间来解决该连续状态空间的 MDP 问题，比如相对值迭代（Relative value iteration, RVI）方法[126]；另外一些工作则采用值函数逼近的方法来解决连续状态空间的 MDP 问题，例如线性函数逼近[127]和神经网络函数逼近[128]。然而，这些工作只能近似地解决问题（4-4），且仍存在计算复杂度高的问题。为了克服这些问题，下面考虑将该调度问题建模成一个 CRMAB 问题。

4.2.4 CRMAB 问题模型

在 CRMAB 问题中，玩家是 BS，动作是信源与信道的组合。在每个时隙开始时，玩家决定需要被激活的信源。在时隙结束时，玩家将得到相应的奖励，即公式（4-1）中的 AoI $\vec{A}(t)=\{A_1(t),\ldots,A_N(t)\}$。令 $d_{n,t}$ 表示信源 n 在时隙 t 的状态，即 $A_n(t)=d_{n,t}$。每个信源根据其选择策略 $u_{nk}(t)$，状态可以分成活跃状态 $\sum_k u_{nk}(t)\neq 0$ 和非活跃状态 $u_{nk}(t)=0,\ \forall k$。下面给出信源 n 的状态转移概率：

· 信源 n 在时隙 t 选择信道 k 但传输失败的状态转移概率为

$$\mathbb{P}\{A_n(t+1)=d_{n,t}\alpha_n(t)+1\mid A_n(t)=d_{n,t},u_{nk}(t)=1\}=1-p_{nk} \tag{4-5}$$

· 信源 n 在时隙 t 选择信道 k 且传输成功的状态转移概率为

$$\mathbb{P}\{A_n(t+1)=1\mid A_n(t)=d_{n,t},u_{nk}(t)=1\}=p_{nk} \tag{4-6}$$

· 信源 n 在时隙 t 没有被调度的状态转移概率为

$$\mathbb{P}\{A_n(t+1)=d_{n,t}\alpha_n(t)+1\mid A_n(t)=d_{n,t},u_{nk}(t)=0\}=1, \tag{4-7}$$

其中，$\{A_n(t+1)\mid A_n(t),u_{nk}(t)\}$ 表示信源 n 选择动作 $u_{nk}(t)$，状态由 $\mathrm{A}_n(t)$ 转移到

$\mathrm{A}_n\left(t+1\right)$的过程。

从式（4-5）至（4-7）可以看到，每个信源的状态不仅与其自身有关，还和它相关的信源状态有关。这将导致系统的状态数目不可数或连续。为了克服这个问题，本节采用平均的 AoI 残余因子$\bar{\alpha}_n \in (0,1]$来代替式上式中的实时 AoI 残余因子$\alpha_n(t)$。其中，平均 AoI 残余因子$\bar{\alpha}_n$通过对T时隙内的随机调度策略取平均得到，即

$$\bar{\alpha}_n = \mathbb{E}\left[\prod_{j\in\mathcal{I}_g^t}(1-e_{jn})\right], \forall n \in \mathcal{N}, \tag{4-8}$$

因此，每一个信源在下一个时隙只有两个可能的状态，即1或者$d_{n,t}\bar{\alpha}_n+1$。这种情况下，网络中的信源可以近似看成独立，且信源上的状态数目变得可数。

下面简要解释该近似的合理性。首先，在基于 AoI 的调度问题中，BS 关心的是信源的 AoI 随时隙变化的趋势，而不是每个信源上 AoI 的具体值[121]，且这种变化趋势本质上反映了信源与信源、信源与不同信道状况之间的相关性。其次，对于给定的调度策略π，网络最终将收敛到一个平稳状态（即网络的平均 AoI 保持稳定）。之后，调度策略$\mathcal{I}_g^t$和 AoI 残余因子$\alpha_n(t)$均呈周期性变化。因此，平均的 AoI 残余因子在网络收敛情况下将是一个与时隙t无关的常数，即$\mathbb{E}[\alpha_n(t)]=\bar{\alpha}_n$。换言之，即使$\bar{\alpha}_n$与时隙$t$无关，其仍能反应出相关信源之间的 AoI 变化趋势。因此，本节考虑采用平均的 AoI 残余因子$\bar{\alpha}_n$来代替式（4-5）至（4-7）中的实时 AoI 残余因子$\alpha_n(t)$。下面采用该平均 AoI 残余因子对 CRMAB 问题进行解耦。

根据以上分析，解耦的 CRMAB 问题可以表示为

$$\begin{aligned}
\min \quad & \lim_{T\to\infty}\sum_{n=1}^{N}\mathbb{E}\left[J_n\right] \\
\text{s.t.} \quad & J_n = \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}\left(A_n(t)-m_k\left(1-u_{nk}(t)\right)\right) \\
& \bar{\alpha}_n = \mathbb{E}\left[\prod_{j\in\mathcal{I}_g^t}(1-e_{jn})\right], \forall n \in \mathcal{N} \\
& \sum_{n=1}^{N} u_{nk}^{\pi}(t) = 1, \forall k \in \mathcal{K} \\
& u_{nk}^{\pi}(t) \in \{0,1\}, \forall k \in \mathcal{K}, \forall t
\end{aligned} \tag{4-9}$$

其中，m_k 是关于信道 k 的一个常数。可以看到，在引入平均 AoI 残余因子 $\bar{\alpha}_n$ 后，信源之间近似独立，问题（4-9）转化成传统的 RMAB 问题。与 MDP 问题相比，RMAB 在计算 WI 时通过将一个 N 维的问题解耦成 N 个独立的一维 MAB 问题来求解，从而有效降低了计算复杂度。此外，RMAB 问题中的还有两个重要概念，即可索引性和 WI。两者的关系为：WI 的闭式表达式存在当且仅当 RMAB 问题是可索引的。因此，在求解 RMAB 问题之前，首先需要证明其可索引性。对于一个可索引 RMAB 问题，通常构建基于 WI 的调度策略来求解，即在每个时隙激活那些具有最大 WI 值的信源来完成其 AoI 的更新。如果一维的 MAB 问题的 WI 只依赖其自身，便将其称为强可分解 WI；然而，为了捕捉信源之间的相关性，本文研究弱可分解 WI，即它的计算不仅依赖于自身，还依赖于与其相关的信源。在下文中将证明强可分解 WI 是弱可分解 WI 的一个特例。所以，将后者称为广义 GWI。

4.3 基于 RMAB 的物联网设备调度策略

本节首先提出一种基于 GWI 的调度策略来解决理想信道模型下的网络平均 AoI 最小化问题，并证明上述 CRMAB 问题的可索引性和推导 GWI 的解析表达式；然后，在非理想信道模型下，提出一种基于 GPWI 的调度策略来解决网络平均 AoI 最小化问题，并证明上述 CRMAB 问题的可索引性和推导 GPWI 的解析表达式。

4.3.1 理想信道下的 GWI 调度策略

根据前面介绍，在理想信道下可以省略信道的索引 k。因此，对于每个信源只有两种可能的动作，即被调度 $u=1$ 和不被调度 $u=0$。为了方便分析，假设符号 $d_{n,t}$ 和 d 可以互换。根据文献[123]，一维的 MAB 问题也可以看作是一个 MDP。下面，令 $V_m(d)$ 表示初始状态为 d 补偿为 m 的可微值函数，其物理意义是该 MDP 可获得的最小预期总奖励。

因此，该一维 MAB 问题的 Bellman 方程为

$$V_m\left(d\right)+J^*=\min\left[V_m\left(d;u=0\right),V_m\left(d;u=1\right)\right] \tag{4-10}$$

其中，J^* 表示最优的平均奖励，$V_m\left(d;u\right)$ 表示在最优策略下信源选择动作 u 获得的

总期望奖励。根据式（4-5）至（4-7）的状态转移概率，上述 Bellman 方程可以重新表示为

$$V_m(d)+J^*=\min\begin{cases}V_m(d;u=0)=d\bar{\alpha}+1-m+V_m(d\bar{\alpha}+1),\\ V_m(d;u=1)=p+(d\bar{\alpha}+1)(1-p)\\ \qquad\qquad +V_m(1)p+V_m(d\bar{\alpha}+1)(1-p)\end{cases}\tag{4-11}$$

其中，初始条件为$V_m(1)=0$。下面分别给出非活跃集合、可索引性和 GWI 的定义。

定义 4.1.（非活跃集合）令$\mathcal{P}(m)$表示在补偿m下使得当前信源不被调度的状态d的集合，它满足以下关系

$$\mathcal{P}(m)=\{d:V_m(d;u=0)\le V_m(d;u=1)\}\tag{4-12}$$

定义 4.2.（可索引性）一个一维 MAB 问题是可索引的，当且仅当补偿m从$-\infty$增加到$+\infty$时，非活跃集合$\mathcal{P}(m)$从空集增加至全集。如果一个 RMAB 问题是可索引的，当且仅当其所有的一维 MAB 问题均是可索引的。

定义 4.3.（GWI）如果一个一维 MAB 问题是可索引的，它的 GWI $W(d,\bar{\alpha})$定义为使得当前信源由不被调用到刚好被调用所需要的最小补偿m，即

$$W(d,\bar{\alpha})=\inf_m\{m:V_m(d;u=0)=V_m(d;u=1)\}\tag{4-13}$$

下面讨论问题（4-9）的可索引性，并推导其 GWI 的解析表达式。根据文献[129]，Bellman 方程（4-11）的最优解是一个门限策略，即

性质 4.1. 对于一维的 MAB 问题，在给定平均残余 AoI 因子的情况下，Bellman 方程(4-11)的最优解可以看作是一个阈值策略，即当$1\le d<D$时，该信源不被调用；当$d\ge D$时，被调用。其中，阈值D的表达式为

$$D=\begin{cases}\dfrac{\bar{\alpha}(1-p)(p-m)+(m-1)}{p\bar{\alpha}(2-p-(1-p)\bar{\alpha})}, & 0<\bar{\alpha}<1\\ -\left(1-m+\dfrac{(1-p)^2}{p}\right), & \bar{\alpha}=1\end{cases}\tag{4-14}$$

于此同时，最优的平均奖励J^*为

$$J^*=\begin{cases}\dfrac{1+Z-pm\tilde{n}}{p\tilde{n}+1}+\dfrac{p(\tilde{n}+p+p\bar{\alpha}-D\bar{\alpha})}{(1+p\tilde{n})(1-\bar{\alpha})}, & 0<\bar{\alpha}<1\\ \dfrac{\left(1-p^2\right)^2-(m^2+2)p^2+2mp}{2p(1-p)}, & \bar{\alpha}=1\end{cases} \tag{4-15}$$

其中，

$$\begin{cases}\tilde{n} &= \log_{\bar{\alpha}}\left(1-D(1-\bar{\alpha})\right)-2,\\ Z &= \dfrac{\bar{\alpha}(1-p)(Dp+1-p)}{1-(1-p)\bar{\alpha}},\end{cases} \tag{4-16}$$

是在$0<\bar{\alpha}<1$情况下的两个中间变量。

证明：Bellman 方程（4-11）可以重写为

$$V_m\left(d\right)=d\bar{\alpha}+1-J^*+V_m(d\bar{\alpha}+1)+\min\begin{cases}-m,\\ -pd\bar{\alpha}-pV_m(d\bar{\alpha}+1),\end{cases} \tag{4-17}$$

其中，式子的上面一部分是关于动作$u=0$的奖励，下面一部分是关于动作$u=1$的奖励。首先，考虑$0<\bar{\alpha}<1$的情况。当$d\geq D$时，可以得到$-pd\bar{\alpha}-pV_m(d\bar{\alpha}+1)<-m$，即

$$V_m(d\bar{\alpha}+1)>\frac{m-pd\bar{\alpha}}{p} \tag{4-18}$$

同时，值函数$V_m(d)$可以表示为

$$V_m(d)=(1-p)V_m(d\bar{\alpha}+1)+(1-p)d\bar{\alpha}+1-J^* \tag{4-19}$$

然后，可以得到以下递归关系

$$\begin{cases}V_m(d)=(1-p)V_m(d\bar{\alpha}+1)+(1-p)d\bar{\alpha}+1-J^*\\ (1-p)V_m(d\bar{\alpha}+1)=(1-p)^2V_m(d\bar{\alpha}^2+\bar{\alpha}+1)+(1-p)^2(d\bar{\alpha}+1)\\ \qquad\qquad +(1-p)(1-J^*)\\ \qquad\qquad\vdots\\ (1-p)^nV_m(d\bar{\alpha}^n+\ldots+1)=(1-p)^{n+1}V_m(d\bar{\alpha}^{n+1}+\ldots+1)+(1-p)^{n+1}\\ \qquad\qquad (d\bar{\alpha}^{n+1}+\ldots+\bar{\alpha}+1)+(1-p)^n(1-J^*)\end{cases} \tag{4-20}$$

通过对上式求和，可以得到

$$\begin{aligned} V_m(d)=(1-p)^{n+1}V_m(d\bar{\alpha}^{n+1}+\ldots+1)+(1-p)d\bar{\alpha}+\ldots+(1-p)^{n+1} \\ (d\bar{\alpha}^{n+1}+\ldots+\bar{\alpha})+(1-J^*)\left(1+(1-p)+\ldots+(1-p)^n\right) \end{aligned} \tag{4-21}$$

因此，当$n\to+\infty$时，有$(1-p)^{n+1}\to 0$和

$$V_m(d)=\frac{\bar{\alpha}(1-p)(dp+1-p)}{p(1-(1-p)\bar{\alpha})}+\frac{1-J^*}{p} \tag{4-22}$$

当$1\le d<D$时，可以得到$-pd\bar{\alpha}-pV_m(d\bar{\alpha}+1)\ge -m$，即

$$V_m(d\bar{\alpha}+1)\le\frac{m-pd\bar{\alpha}}{p} \tag{4-23}$$

同时，值函数$V_m(d)$可以表示为

$$V_m(d)=V_m(d\bar{\alpha}+1)-m+d\bar{\alpha}+1-J^* \tag{4-24}$$

将$d=\left(D-1\right)/\bar{\alpha}$代入上式，得到

$$V_m(D)=V_m\left(\frac{D}{\bar{\alpha}}-\frac{1}{\bar{\alpha}}\right)-D+(J^*+m) \tag{4-25}$$

然后，可以得到以下递归关系

$$\begin{cases} V_m(D)=V_m\left(\dfrac{D}{\bar{\alpha}}-\dfrac{1}{\bar{\alpha}}\right)-D+(J^*+m) \\ V_m\left(\dfrac{D}{\bar{\alpha}}-\dfrac{1}{\bar{\alpha}}\right)=V_m\left(\dfrac{D}{\bar{\alpha}^2}-\dfrac{1}{\bar{\alpha}^2}-\dfrac{1}{\bar{\alpha}}\right)-\left(\dfrac{D}{\bar{\alpha}}-\dfrac{1}{\bar{\alpha}}\right)+(J^*+m) \\ \vdots \\ V_m\left(\dfrac{D}{\bar{\alpha}^n}-\dfrac{1}{\bar{\alpha}^n}-\ldots-\dfrac{1}{\bar{\alpha}}\right)=V_m\left(\dfrac{D}{\bar{\alpha}^{n+1}}-\dfrac{1}{\bar{\alpha}^{n+1}}-\ldots-\dfrac{1}{\bar{\alpha}}\right) \\ \qquad\qquad -\left(\dfrac{D}{\bar{\alpha}^n}-\dfrac{1}{\bar{\alpha}^n}-\ldots-\dfrac{1}{\bar{\alpha}}\right)+(J^*+m) \end{cases} \tag{4-26}$$

通过对上式求和，可以得到

$$\begin{aligned} V_m\left(\frac{D}{\bar{\alpha}^{n+1}}-\frac{1-\bar{\alpha}^{n+1}}{\bar{\alpha}^{n+1}-\bar{\alpha}^{n+2}}\right)=V_m(D)+n\left(\frac{1}{1-\bar{\alpha}}-J^*-m\right) \\ +\frac{\bar{\alpha}^{n+1}-1}{\bar{\alpha}^{n+1}-\bar{\alpha}^n}D+\frac{\bar{\alpha}^n-1}{\bar{\alpha}^n(1-\bar{\alpha})^2} \end{aligned} \tag{4-27}$$

现在已经有了关于阈值D、最优平均奖励J^*和补偿m的值函数表达式。为了求出

这三个变量，还需要以下两个条件：$V_m(1)=0$ 和 $V_m(D+J^*)=(m-pD\bar{\alpha})/p$。其中，第一个是初始条件恒成立；第二个是因为存在一个最优状态 $(D+J^*)$，使得不等式（4-18）和（4-23）成立，即

$$V_m(D)\le\frac{m-pD\bar{\alpha}}{p}<V_m(D\bar{\alpha}+1) \tag{4-28}$$

因此，可以认为存在一个 $J^*\in[0,1)$ 使得 $V_m(D+J^*)=(m-pD\bar{\alpha})/p$。结合以上两个条件，可以得到

$$\begin{cases}\dfrac{\bar{\alpha}(1-p)(1-p+(D+J^*p))}{1-(1-p)\bar{\alpha}}=J^*+m-pD\bar{\alpha}-1\\ \dfrac{-m+Dp\bar{\alpha}}{p}+\dfrac{(1-p)\bar{\alpha}J^*}{1-(1-p)\bar{\alpha}}=\tilde{n}(\dfrac{1}{1-\bar{\alpha}}-J^*-m)+\dfrac{\bar{\alpha}^{\tilde{n}+1}-1}{\bar{\alpha}^{\tilde{n}+1}-\bar{\alpha}^{\tilde{n}}}D+\dfrac{\bar{\alpha}^{\tilde{n}}-1}{\bar{\alpha}^{\tilde{n}}(1-\bar{\alpha})^2}\end{cases} \tag{4-29}$$

其中，$\tilde{n}=\log_{\bar{\alpha}}\left(1-D(1-\bar{\alpha})\right)-2$，通过求解以下等式得到

$$\frac{D}{\bar{\alpha}^{n+1}}-\frac{1-\bar{\alpha}^{n+1}}{\bar{\alpha}^{n+1}-\bar{\alpha}^{n+2}}=1 \tag{4-30}$$

联立公式（4-29）得到

$$D=\frac{\bar{\alpha}(1-p)(p-m-J^*p-J^*)+(J^*+m-1)}{p\bar{\alpha}(2-p-(1-p)\bar{\alpha})} \tag{4-31}$$

由于

$$\frac{\partial D}{\partial J^*}=\frac{1-(1-p^2)\bar{\alpha}}{p\bar{\alpha}(2-p-(1-p)\bar{\alpha})}>0 \tag{4-32}$$

$D(J^*)$ 在区间 [0, 1) 内单调递增。因此，最优值可以在区间的端点取得。也就是说，令 $J^*=0$，可以得到阈值 D 的表达式

$$D=\frac{\bar{\alpha}(1-p)(p-m)+(m-1)}{p\bar{\alpha}(2-p-(1-p)\bar{\alpha})} \tag{4-33}$$

此外，根据式（4-29）和（4-33），可以得到最优平均值 J^* 的表达式

$$J^*=\frac{1+Z+pm\tilde{n}}{p\tilde{n}+1}+\frac{p(\tilde{n}+p+p\bar{\alpha}-D\bar{\alpha})}{(1+p\tilde{n})(1-\bar{\alpha})} \tag{4-34}$$

其中，

$$Z=\frac{\bar{\alpha}(1-p)(DP+1-p)}{1-(1-p)\bar{\alpha}} \tag{4-35}$$

最后，证明式（4-33）是一个阈值策略，即需要满足以下条件

$$V_m\left(\frac{D}{\bar{\alpha}^{1+n^-}}-\frac{1-\bar{\alpha}^{1+n^-}}{\bar{\alpha}^{1+n^-}-\bar{\alpha}^{2+n^-}}\right)\le\frac{m-pD\bar{\alpha}}{p}<V_m(d') \tag{4-36}$$

其中，$d'\in[D,+\infty)$和$n^-\in\{0,1,2,\ldots\}$。因此，下面只需证明D满足以上不等式。根据（4-24）和（4-27），可以发现值函数是关于d'和n^-的单调递增函数。因此，可以证明式（4-33）是一个阈值策略，即当$1\le d<D$时，该信源不被调用；当$d\ge D$时，被调用。

此外，对于$\bar{\alpha}=1$的情况，同理可证。 □

注 1. 从性质 4.1 可以看出，在$0<\bar{\alpha}<1$的情况下，阈值D是补偿m的线性函数，即当m从$-\infty$到$+\infty$变化时，阈值D线性增加。因此，存在一个最优补偿m^*，使得非活跃集合$\mathcal{P}(m)=\varnothing$，即$D(m^*)=1$。这表明当$m$从$-\infty$增加到$+\infty$时，非活跃集合$\mathcal{P}(m)$单调地从空集增加到整个状态空间。根据定义 4.2，可以证明问题（4-9）是可索引的，因为每个一维的 MAB 问题都是可索引的。

注 2. 根据定义 4.3，阈值D必须是$d\bar{\alpha}+1$。将$D=d\bar{\alpha}+1$代入式（4-14）中，得到

$$m(d,\bar{\alpha})=\begin{cases}\left(p\bar{\alpha}(d+1)+\dfrac{1+pd\bar{\alpha}(1-p)}{1-\bar{\alpha}(1-p)}\right), & 0<\bar{\alpha}<1\\ \dfrac{p^2+dp+1}{p}, & \bar{\alpha}=1\end{cases} \tag{4-37}$$

其中，$m(d,\bar{\alpha})=W\left(d,\bar{\alpha}\right)$表示 GWI 的闭式表达式。

下面给出两种计算平均残余 AoI 因子$\bar{\alpha}_n$的方法，即实时的残余 AoI 因子和估计的残余 AoI 因子。实时的残余 AoI 因子是指在中心式系统中，BS 知道的网络中信源之间的相关矩阵$\mathcal{E}$的先验信息，此时平均残余 AoI 因子$\bar{\alpha}_n$由式（4-8）给出。

估计的残余 AoI 因子是指在分布式系统中相关矩阵$\mathcal{E}$未知，平均残余 AoI 因子$\bar{\alpha}_n$需要从历史数据中估计出来。根据（4-1），AoI 演化公式可以写成斜率为$\bar{\alpha}_n$的线性方程，即

$$A_n(t+1)=A_n(t)\bar{\alpha}_n+1 \tag{4-38}$$

为了估计 $\bar{\alpha}_n$，令 $H_y = A_n(t+1)-1$ 和 $H_x = A_n(t)-1$，对于 $t=\{1,2,\ldots,t\}$。利用最小二乘（Least Square Estimation, LSE）方法，得到

$$\hat{\alpha}_n = \frac{\sum_{i=1}^{t}\left(H_{x_i}-\bar{H}_x\right)\left(H_{y_i}-\bar{H}_y\right)}{\sum_{i=1}^{t}\left(H_{x_i}-\bar{H}_x\right)^2} \tag{4-39}$$

其中，

$$\bar{H}_x = \frac{1}{t}\sum_{i=1}^{t}\left(A_n(i)-1\right) \tag{4-40}$$

和

$$\bar{H}_y = \frac{1}{t}\sum_{i=2}^{t+1}\left(A_n(i)-1\right) \tag{4-41}$$

需要注意的是，奖励 $A_n(t+1)$ 需要利用线性预测滤波（linear prediction filter coefficients，LPC）方法从历史奖励中预测得到。

算法 4.1 基于 GWI 的调度策略

步骤 1：	初始化参数： $K, N, T, p_n, \mathcal{E}, A_n(1)=1$
步骤 2:	执行 $t=1,2,...,T$ 次以下步骤
步骤 3:	根据式（4-37）计算每一个信源的 GWI
步骤 4:	激活前 K 个关于 $W(d,\bar{\alpha})$ 的最大值的信源
步骤 5:	根据传输结果更新每一个信源的 AoI
步骤 6:	利用（4-8）或（4-39）计算平均残余 AoI 因子 $\bar{\alpha}$

最后，根据上面的分析，给出理想信道下基于 GWI 方法的调度策略来求解 CRMAB 问题（4-9），如算法 4.1 所示。可以看到，在每个时隙开始时，BS 根据公式（4-37）计算每个信源上的 GWI。接着，BS 激活 GWI 中前 K 个最大值的信源来更新它们的状态。然后，BS 根据当前时隙被调度的信源的传输结果更新每个信源的 AoI。最后，使用式（4-8）或（4-39）计算平均残余 AoI 因子 $\bar{\alpha}_n$。算法 4.1 重复以上步骤，直到时间 T。

4.3.2 非理想信道下的 GPWI 调度策略

在非理想信道下，问题（4-9）中一维的 MAB 问题同样可以看作是一个 MDP 问题。但每个信源有 $K+1$ 个可能的动作，即 $u=\{0,1,2,\ldots,K\}$，其中，$u=0$ 表示信源没有被调度。因此，一维的 MAB 问题的 Bellman 方程可以表示为

$$\begin{aligned} V_m(d)+J^* &= \min_{u\in\{0,1,\ldots,K\}} V_m(d,u) \\ &= \min_{u\in\{0,1,\ldots,K\}} \left\{ r_u(d)+\sum_{d'} \mathbb{P}_{d'd}(u) V_{\vec{m}}(d') \right\} \end{aligned} \tag{4-42}$$

其中，$\mathbb{P}_{d'd}(u)$ 表示信源在动作 u 时，状态由 d 转移到 d' 的概率；$r_u(d)$ 表示信源初始状态为 d 和采用动作 u 获得的累积奖励和补偿 m 之和；$\vec{m}=\{m_1,m_2,\ldots,m_K\}$ 是关于 K 个信道的补偿向量。与理想信道模型相比，信源被调度在不同的信道会产生不同的奖励。根据（4-42），选择信道 k 的决策不仅取决于该信道，还与其相关的信道的补偿有关。因此，无法用单一的阈值将一个信源上的状态空间划分为活跃集合和非活跃状态集合，即上述 Bellman 方程可以改写为

$$\begin{aligned} V_{\vec{m}}(d)+J^* &= \min_{u\in\{0,1,\ldots,K\}} V_{\vec{m}}(d,u) \\ &= \min_{u\in\{1,\ldots,K\}} \min_{k} \left[V_{m_k}\left(d;u=0\right), V_{m_k}\left(d;u=k\right) \right] \end{aligned} \tag{4-43}$$

其中，$V_{m_k}\left(d;u=0\right)$ 是关于补偿 m_k 的值函数。

性质 4.2. 在非理想信道下，Bellman 方程可以分解成 K 个独立的 MAB 问题，每一个子问题的动作空间为 $u=\{0,1\}$。

证明：在非理想信道下，各个信道经历的衰落不同，但各信道之间相互独立。因此，上述 Bellman 方程可以写成

$$V_{\vec{m}}(d)+J^* = \min_{u\in\{1,\ldots,K\}} \begin{cases} \min\left[V_{m_1}\left(d;u=0\right), V_{m_1}\left(d;u=1\right) \right] \\ \min\left[V_{m_2}\left(d;u=0\right), V_{m_2}\left(d;u=2\right) \right] \\ \vdots \\ \min\left[V_{m_K}\left(d;u=0\right), V_{m_K}\left(d;u=K\right) \right] \end{cases} \tag{4-44}$$

其中，$\vec{m}=\{m_1,m_2,\ldots,m_K\}$。令 $\vec{m}_{-k}$ 表示除了信道 k 的补偿向量和 $\vec{m}'=[m_k',\vec{m}_{-k}]$ 表示

固定向量 $\vec{m}_{-k}$ 中的值，但改变 m_k' 的新向量，则上式可以变成

$$\min_{u\in\{1,\ldots,K\}}\begin{cases} V_{\vec{m}'}(d)+J^*=\min\left[V_{m_1}(d;u=0),V_{m_1}(d;u=1)\right] \\ V_{\vec{m}'}(d)+J^*=\min\left[V_{m_2}(d;u=0),V_{m_2}(d;u=2)\right] \\ \vdots \\ V_{\vec{m}'}(d)+J^*=\min\left[V_{m_K}(d;u=0),V_{m_K}(d;u=K)\right] \end{cases} \tag{4-45}$$

因此，Bellman 方程可以分解成 K 个独立的 MAB 问题，且每个子问题的动作空间为 $u=\{0,1\}$。 □

因此，上述 Bellman 方程可以和理想信道情况下一样求解。在推导本节主要结果前，先给出非活跃集合、部分可索引性和 GPWI 的定义。

定义 4.4.（非活跃集合）给定补偿向量 $\vec{m}$，非活跃集合 $\mathcal{P}_k(m)$ 表示在补偿 $\vec{m}$ 下使得当前信源不被调度在信道 k 的状态 d 的集合，它满足以下关系

$$\mathcal{P}_k(\vec{m})=\{d:V_{m_k}(d)>\min_{u\neq k}V_{m_u}(d)\} \tag{4-46}$$

定义 4.5.（可索引性）令 $\vec{m}_{-k}$ 表示除了信道 k 的补偿向量和 $\vec{m}'=[m_k',\vec{m}_{-k}]$ 表示固定向量 $\vec{m}_{-k}$ 中的所有值,但改变 m_k' 的新向量。然后,一个一维 MAB 问题是可索引的，当且仅当补偿 m_k' 从 $-\infty$ 增加到 $+\infty$ 时，非活跃集合 $\mathcal{P}_k(\vec{m}')$ 从空集增加至全状态空间。如果一个 RMAB 问题是可索引的，当且仅当其所有的 $N\times K$ 个一维 MAB 问题均是可索引的。

定义 4.6.（GPWI）给定补偿向量 $\vec{m}$ 和 $\vec{m}_{-k}$，如果一个一维 MAB 问题是部分可索引的，它的 GPWI $G_k(d,\bar{\alpha})$ 表示使得当前信源由不被调用到刚好被调用在信道 k 所需要的最小补偿 m_k，即

$$G_k(d,\bar{\alpha})=\inf_{m_k'}\left\{m_k':V_{m_k'}(d;u=0)=V_{m_k'}(d;u=k)\right\} \tag{4-47}$$

从上面定义可以看到，GPWI 的定义与信道相关。因此，对于一个解耦的 CRMAB 问题，GPWI 是一个 $N\times K$ 的矩阵。与理想信道情况下类似，下面证明和分析问题（4-9）在非理想信道下的相关性质。

性质 4.3. 在给定补偿向量$\vec{m}$和$\vec{m}_{-k}$，对于一维的 MAB 问题，最优的调度策略是基于 Bellman 方程的阈值策略，即当$1 \le d < D_k$时，该信源不被调用在信道k；当$d \ge D_k$时，被调用在信道k；其中，阈值D_k的表达式为

$$D_k = \begin{cases} \dfrac{\bar{\alpha}(1-p_k)(p_k - m_k) + (m_k - 1)}{p_k \bar{\alpha}(2 - p_k - (1-p_k)\bar{\alpha})}, & 0 < \bar{\alpha} < 1 \\ -\left(1 - m_k + \dfrac{(1-p_k)^2}{p_k}\right), & \bar{\alpha} = 1 \end{cases} \tag{4-48}$$

同时，最优的平均奖励J_k^*为

$$J_k^* = \begin{cases} \dfrac{1 + Z_k - p_k m_k \tilde{n}_k}{p_k \tilde{n}_k + 1} + \dfrac{p_k(\tilde{n}_n + p_k + p_k\bar{\alpha} - D_k\bar{\alpha})}{(1 + p_k\tilde{n})(1-\bar{\alpha})}, & 0 < \bar{\alpha} < 1 \\ \dfrac{\left(1 - p_k^2\right)^2 - (m_k^2 + 2)p_k^2 + 2m_k p_k}{2p_k(1-p_k)}, & \bar{\alpha} = 1 \end{cases} \tag{4-49}$$

其中

$$\begin{cases} \tilde{n}_k &= \log_{\bar{\alpha}}\left(1 - D_k(1-\bar{\alpha})\right) - 2 \\ Z_k &= \dfrac{\bar{\alpha}(1-p_k)(D_k p_k + 1 - p_k)}{1 - (1-p_k)\bar{\alpha}} \end{cases} \tag{4-50}$$

证明：根据性质 4.2，每一个信源的 Bellman 方程可以分解成K个独立的 MAB 问题，且每个子问题的动作空间为$u = \{0,1\}$。因此，可以采用性质 4.1 中相同的方法分析每一个独立的 MAB 问题的阈值D_k。最后，根据性质 4.1 中的结果，分别考虑单个信道的情况，得到阈值D_k和最优平均奖励J_k^*的表达式。 □

注 3. 同理可证，在非理想信道下，CRMAB 问题是部分可索引的，因为所有一维 MAB 问题在不同信道都是可索引的。

注 4. 同理可证，在非理想信道下，通过将$D = d\bar{\alpha} + 1$代入式（4-48），可得

$$m_k(d, \bar{\alpha}) = \begin{cases} \left(p_k\bar{\alpha}(d+1) + \dfrac{1 + p_k d\bar{\alpha}(1-p_k)}{1 - \bar{\alpha}(1-p_k)}\right), & 0 < \bar{\alpha} < 1 \\ \dfrac{p_k^2 + dp_k + 1}{p_k}, & \bar{\alpha} = 1 \end{cases} \tag{4-51}$$

结合定义 4.6，第n个一维 MAB 问题关于信道k的 GPWI 表达式为

$$G_{nk}(d,\bar{\alpha})=\begin{cases}\left(p_{nk}\bar{\alpha}(d+1)+\dfrac{1+p_{nk}d\bar{\alpha}(1-p_{nk})}{1-\bar{\alpha}(1-p_{nk})}\right), & 0<\bar{\alpha}<1\\ \dfrac{p_{nk}^2+dp_{nk}+1}{p_{nk}}, & \bar{\alpha}=1\end{cases} \tag{4-52}$$

根据得到的 $N\times K$ 个 GPWI 表示式，解耦后的 CRMAB 问题（4-9），可以描述为一个最大权重匹配（maximum weighted matching，MWM）问题，即

$$\begin{aligned}\max_{u_{nk}\in\{0,1\}}\quad & \sum_{n=1}^{N}\sum_{k=1}^{K}G_{nk}u_{nk}\\ \text{s.t.}\quad & \sum_{n=1}^{N}u_{nk}=1,\ \forall k\in\mathcal{K}\\ & \sum_{k=1}^{K}u_{nk}\le 1,\ \forall n\in\mathcal{N}\end{aligned} \tag{4-53}$$

因此，基于 GPWI 的调度策略由上述解给出，即 $u_{nk}=1$ 表示信源 n 应被调度至信道 k；否则，$u_{nk}=0$。最后，基于 GPWI 的调度策略由算法 4.2 给出。

算法 4.2 基于 GPWI 的调度策略

步骤 1：	初始化参数： $K,N,T,p_n,\mathcal{E},A_n(1)=1$
步骤 2:	执行 $t=1,2,...,T$ 次以下步骤
步骤 3:	根据式（4-52）计算每一个信源关于信道的 GPWI
步骤 4:	求解 MWM 问题得到决策变量 u_{nk}
步骤 5:	根据决策变量 u_{nk} 调度信源进行数据传输
步骤 6:	根据传输结果更新每一个信源的 AoI
步骤 7:	利用（4-8）或（4-39）计算平均残余 AoI 因子 $\bar{\alpha}$

4.4 理论分析

本节分别给出理想和非理想信道情况下所提策略的数值性能下界。该性能下界通过求解解耦后的 CRMAB 问题的拉格朗日松弛问题得到。具体地，在 CRMAB 问题中，每个时隙应被调度的信源数量严格限制为 K 个。然而，在拉格朗日松弛问题中，只需要在时间 T 内平均的信源调度数目等于 K 即可。这种情况下，拉格朗

日松弛问题的解将优于 CRMAB 问题的解，从而为所提策略提供一个理论性能下界。

根据前面分析，原问题可以重写为

$$\begin{aligned}\min_{d}\quad & \frac{1}{T}\sum_{t=1}^{T}\sum_{n=1}^{N}\mathbb{E}\left[A_n(t)\right]\\ \text{s.t.}\quad & \bar{\alpha}_n=\mathbb{E}\left[\prod_{j\in\mathcal{I}_g^t}(1-e_{jn})\right],\forall n\in\mathcal{N},\forall t\\ & \sum_{n=1}^{N}u_{nk}(t)=1,\forall k\in\mathcal{K}\\ & u_{nk}(t)\in\{0,1\},\forall t\end{aligned}\tag{4-54}$$

其中，第二个约束条件等价于$\sum_{n=1}^{N}\left(1-u_{nk}(t)\right)=N-1$。因此，该松弛问题可以表示为

$$\begin{aligned}\min_{d}\quad & \frac{1}{T}\sum_{t=1}^{T}\sum_{n=1}^{N}\mathbb{E}\left[A_n(t)\right]\\ \text{s.t.}\quad & \bar{\alpha}_n=\mathbb{E}\left[\prod_{j\in\mathcal{I}_g^t}(1-e_{jn})\right],\forall n\in\mathcal{N},\forall t\\ & \frac{1}{T}\sum_{t=1}^{T}\sum_{n=1}^{N}\mathbb{E}\left[\left(1-u_{nk}(t)\right)\right]=N-1,\forall k\in\mathcal{K}\\ & u_{nk}(t)\in\{0,1\},\forall t\end{aligned}\tag{4-55}$$

接着，上式的拉格朗日函数为

$$L(d,\vec{m})=\inf_{d}\sum_{n=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}\left[A_n(t)\right]-\sum_{k=1}^{K}m_k\left(1-u_{nk}(t)\right)\right)\right)+\sum_{k=1}^{K}m_k(N-1)\tag{4-56}$$

其中，$\vec{m}=\{m_1,m_2,\ldots,m_K\}$表示拉格朗日乘子向量。

实际上，由于目标函数的定义域是整个系统状态空间，很难直接求解该拉格朗日松弛问题。但是，通过观察可以发现，目标函数中的第一项是解耦后的 CRMAB 问题的目标函数。因此，可以将（4-56）的第一项替换为$\sum_{n=1}^{N}J_n^*$，其中J_n^*的表达式由（4-15）和（4-49）给出，分别对应于理想信道和非理想信道的情况。接着，构造出松弛问题（4-55）的拉格朗日对偶问题。根据凸优化理论可知，通过求解这个对偶问题，可以为松弛问题（4-55）提供一个性能下界[102]。

首先，考虑理想信道的情况，每一个信源只有两个动作。问题（4-55）的拉格朗日对偶问题可以表示成

$$\begin{aligned}\min_{m}\quad & \Gamma(m)\\ \text{s.t.}\quad & \bar{\alpha}_n=\mathbb{E}\left[\prod_{j\in\mathcal{I}_g^t}(1-e_{jn})\right],\forall n\in\mathcal{N},\forall t\\ & \Gamma(m)=\sum_{n=1}^{N}J_n^*+m(N-K)\\ & u_n(t)\in\{0,1\},\forall t\end{aligned}\tag{4-57}$$

引理 4.1. 根据表达式（4-15），J_n^* 是关于乘子 m 的凹函数。

证明: 首先，考虑 $0<\bar{\alpha}<1$ 的情况。从式（4-15）容易可以看到，J_n^* 是关于 m 的线性函数。因此，可得到 J_n^* 是关于 m 的二阶导数为零，即其为凹函数。其次，考虑 $\bar{\alpha}=1$ 的情况。同样，利用式（4-15），求 J_n^* 关于 m 二阶导数，即

$$\frac{\partial^2 J^*(m)}{\partial m^2}=\frac{-p}{1-p}\tag{4-58}$$

因为 $0<p<1$，可以得到 $\partial^2 J^*(m)/\partial m^2<0$，即其为凹函数。综上所述，可以证明 $J^*(m)$ 在两种情况下都是关于 m 的凹函数。 □

根据引理 4.1，令目标函数（4-57）关于 m 的导数等于零，得到

$$m^*=\begin{cases}\displaystyle\sum_{n=1}^{N}\frac{\partial J_n^*}{\partial m}+(N-K)=0, & 0<\bar{\alpha}<1\\ \dfrac{(N-K)\displaystyle\sum_{n=1}^{N}(1-p_n)+N}{\displaystyle\sum_{n=1}^{N}p_n}, & \bar{\alpha}=1\end{cases}\tag{4-59}$$

其中，$0<\bar{\alpha}<1$ 情况下的 m^* 可以利用牛顿法得到数值解。

其次，考虑非理想信道的情况，每一个信源有 $K+1$ 种动作。根据性质 4.2，每一个信源对应的 Bellman 方程可以分解成 K 个一维的 MAB 问题。因此，拉格朗日函数（4-56）可以转换成

$$\begin{aligned} L(d,\vec{m}) &= \sum_{n=1}^{N}\sum_{k=1}^{K} J_{nk}^{*} + \sum_{k=1}^{K} m_k (N-1) \\ &= \sum_{k=1}^{K}\left(\sum_{n=1}^{N} J_{nk}^{*} + m_k (N-1) \right) \end{aligned} \tag{4-60}$$

其中，J_{nk}^{*} 的表达式由（4-49）给出。由于 K 个信道是相互独立的，其等价于求解以下拉格朗日对偶问题

$$\begin{aligned} \min_{m_k} \quad & \Gamma(m_k) \\ \text{s.t.} \quad & \bar{\alpha}_n = \mathbb{E}\left[\prod_{j\in\mathcal{I}_g^t} (1-e_{jn}) \right], \forall n \in \mathcal{N}, \forall t \\ & \Gamma(m_k) = \sum_{n=1}^{N} J_{nk}^{*} + m_k (N-K) \\ & u_{nk}(t) \in \{0,1\}, \forall t \end{aligned} \tag{4-61}$$

同理，通过求解上式，可以得到最优拉格朗日乘子 m_k^{*} 的表达式为

$$m_k^{*} = \begin{cases} \sum_{n=1}^{N} \dfrac{\partial J_{n,k}^{*}}{\partial m_k} + N - 1 = 0, & 0 < \bar{\alpha} < 1 \\ \dfrac{(N-1)\sum_{n=1}^{N}(1-p_{nk}) + N}{\sum_{n=1}^{N} p_{nk}}, & \bar{\alpha} = 1 \end{cases} \tag{4-62}$$

通过求解 K 个一维的 MAB 问题，进而得到最优拉格朗日乘子的向量为 $\vec{m}^{*} = \{m_1^{*}, m_2^{*}, \ldots, m_K^{*}\}$。

实际上，该松弛问题揭示了补偿 m 作为拉格朗日乘子的作用，以及所提策略在解耦后的 CRMAB 问题的渐近最优特性。一方面，在松弛约束下，其作用是通过激活当前 GWI 或 GPWI 大于 m^{*} 的信源进行数据传输；另一方面，该策略可以使得松弛约束得到满足，同时其对偶问题达到最大值。根据优化理论，拉格朗日对偶问题的解进一步为松弛问题提供了一个的理论性能下界。综上所述，算法 4.3 给出了基于 GWI 和 GPWI 调度策略的数值性能下界。

算法 4.3 计算理想和非理想信道下所提调度策略的性能下界

步骤 1： 初始化参数： $K, N, T, p_n, \mathcal{E}, A_n(1)=1$

步骤 2: 执行 $t=1,2,...,T$ 次以下步骤

步骤 3: 根据（4-37）和（4-52）计算每一个信源的 GWI 和 GPWI

步骤 4: 利用（4-59）和（4-62）计算得到 m^* 和 m_k^*

步骤 5: 调度 GWI 和 GPWI 高于 m^* 和 m_k^* 的信源在相应的信道中传输

步骤 6: 根据传输结果更新每一个信源的 AoI

步骤 7: 利用（4-8）或（4-39）计算平均残余 AoI 因子 $\bar{\alpha}$

4.5 仿真结果

本节通过数值仿真来验证所提调度策略的有效性。其中，仿真参数主要参照 3GPP 标准[130]，且所有结果均由 10^3 次蒙特卡洛仿真得到。

4.5.1 参数设置与对比算法

仿真中采用室内混合模型[130]来描述信道的直视分量（Line-of-Sight，LoS）分量，则信源 n 的传输成功概率可以表示为

$$p_n=\begin{cases}1, & L_{n,n_0}\le 1.2\text{m},\\ \exp\left(-\dfrac{L_{n,n_0}-1.2}{4.7}\right), & 1.2\text{m}<L_{n,n_0}\le 6.5\text{m},\\ \exp\left(-\dfrac{L_{n,n_0}-6.5}{32.6}\right)\cdot 0.32, & 6.5\text{m}\le L_{n,n_0},\end{cases} \tag{4-63}$$

其中，L_{n,n_0} 表示信源 n 到基站 n_0 之间的欧式距离。然而，对于非理想信道情况，需要综合考虑信道的（Non-Line-of-Sight, NLoS）分量。令 ξ_{nk} 表示信源 n 在信道 k 上传输时，信道的 NLoS 分量。因此，信源 n 在信道 k 上的传输成功概率为 $p_{nk}=\xi_{nk}p_n$，其中，ξ_{nk} 服从 (0, 1) 之间的均匀分布。

考虑空间相关性来描述信源之间的相关性[①]，即两信源之间的相关性与其欧式距离近似成反比[124]。因此，相关矩阵$\mathcal{E}$中的元素e_{ij}可以表示为

$$e_{ij} = \omega_i \exp\left(-\kappa L_{i,j}\right) \tag{4-64}$$

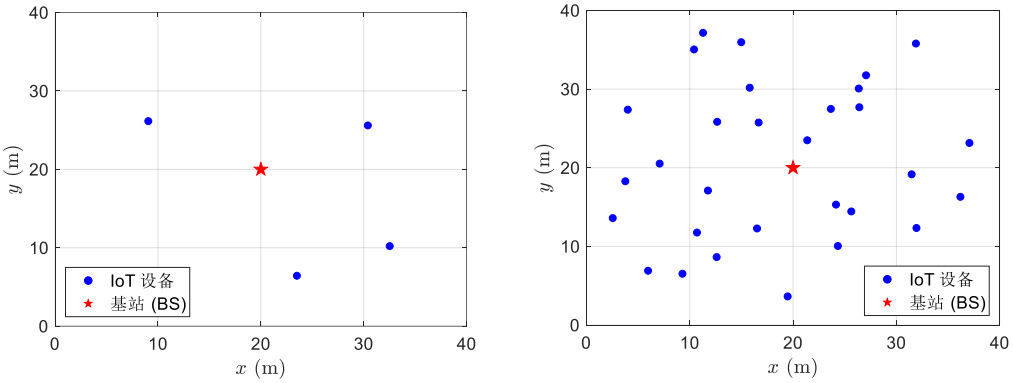
其中，$\omega_i \in (0,1)$是信源i的权重因子，体现信源i在整个网络中的重要性；$L_{i,j}$表示信源i和信源j之间的欧式距离；κ是一个相关因子，控制两信源之间的相关强度。在仿真中，设定$\kappa = 0.05$。值得注意的是，由于$\omega_i \in (0,1)$是一个随机变量，所以相关矩阵$\mathcal{E}$是非对称的。

在仿真中，本节对比了不同的基于 AoI 的调度策略，比如贪婪调度策略、随机调度策略、最大权重匹配策略、理论性能下界。其中，贪婪调度策略是指：在每个时隙t，BS 选择 AoI 值$A_n(t)$中最大的K个信源进行状态更新；随机调度策略是指：在每个时隙t，BS 选择概率值$\eta_n \Big/ \sum_{i=1}^{N} \eta_i$中最大的$K$个信源进行状态更新，其中，$\eta_n$是一个与信源$n$相关的参数；根据文献[121]，最大权重调度策略是指：在每个时隙t，BS 选择概率值$p_n A_n(t)(A_n(t)+2)$最大的K个信源进行状态更新，其中，p_n表示信源n传输成功的概率；最后，理论性能下界由 4.4 节中的算法 4.3 得到。

4.5.2 数值结果分析

下面根据以上参数设置来验证所提算法与对比算法的性能。首先，考虑静态网络对算法性能的影响。本节主要考虑两种静态网络场景，如图 4.3 所示。在一个$(40\times 40)\,\mathrm{m}$的方形区域内，网络场景一和网络场景二中分别存在$N=4$和$N=30$个信源或 IoT 设备。这些 IoT 设备服从密度为$\lambda = N/(\pi R^2)$的齐次泊松分布过程，其中，$R = 20\,\mathrm{m}$是该方形区域内最大的圆的半径。在该静态网络场景中，假设设备的位置产生后不在发生变化。

[①] 由于本节主要目的是验证所提算法的性能，因此没有考虑不同信源样本之间的相关性。



（a）网络场景一　　（b）网络场景二

图 4.3 两个静态 IoT 网络场景图

图 4.4 是在网络场景一和不同残余 AoI 因子条件下网络平均 AoI 随时隙变化的曲线。其中，实时的残余 AoI 因子由式（4-8）给出，估计的残余 AoI 因子由式（4-39）给出；总信道数为$K=2$，总时隙数为$T=200$，网络平均 AoI 的计算方法为$\left(\sum_{i=1}^{t}\sum_{n=1}^{N}A_n(i)\right)/t$。从图中可以看到，在理想和非理想信道情况下，基于 GWI 和 GPWI 的调度策略在实时的残余 AoI 因子下的性能优于估计的残余 AoI 因子下的性能。这是因为实时$\bar{\alpha}$具有相关矩阵$\mathcal{E}$和每次动作的先验信息。从图中还可以看到，基于 GWI 的调度策略优于基于 GPWI 的策略，其原因是非理想信道是由 NLoS 和 LoS 分量联合相乘得到（注意这两个分量都小于 1）；而理想信道只考虑了信道的 LoS 分量。此外，图 4.4 还表明，即使相关矩阵$\mathcal{E}$的先验信息未知，基于 GWI 和 GPWI 的调度策略仍可以通过估计的$\bar{\alpha}$，使得网络的平均 AoI 收敛。

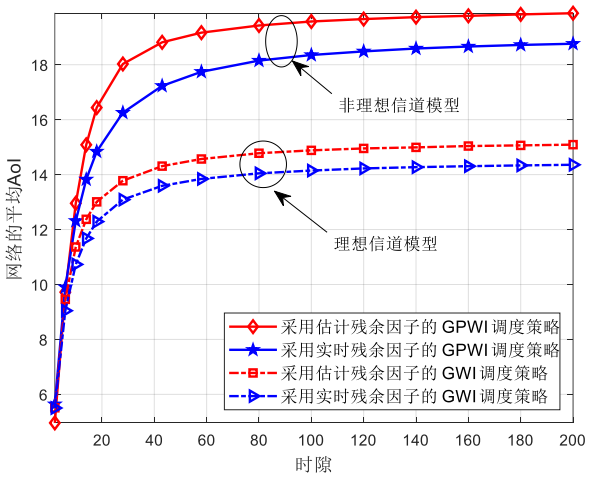
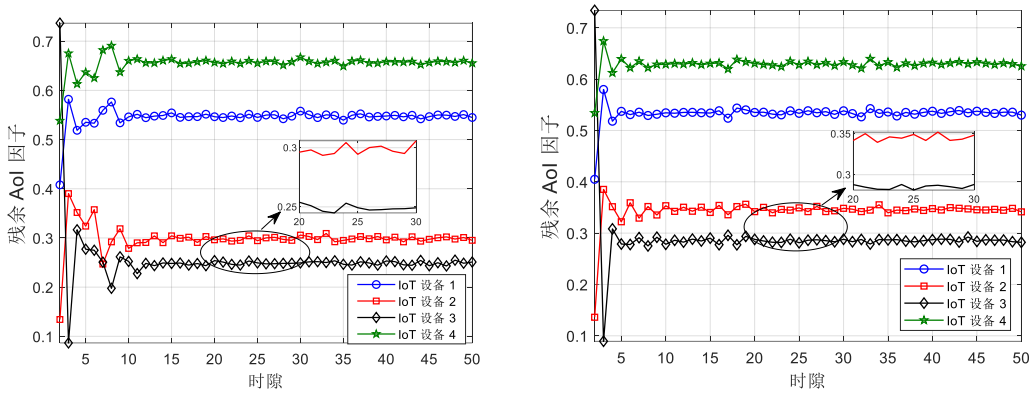


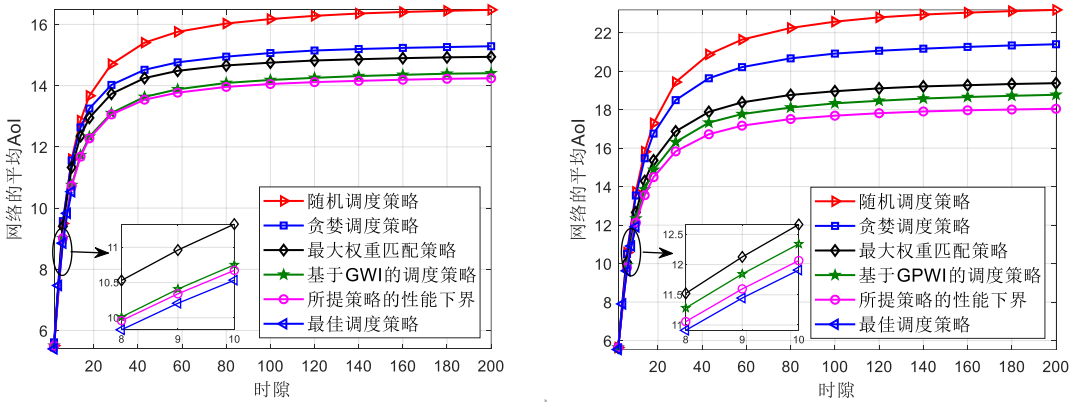
图 4.4 所提调度策略在图 4.3a 网络场景和不同残余 AoI 因子下的性能曲线



（a）利用估计的残余 AoI 因子 $\hat{\alpha}$　　（b）利用实时的残余 AoI 因子 $\bar{\alpha}$

图 4.5 在图 4.3a 网络场景下运行 GWI 调度策略的不同信源的残余 AoI 因子曲线

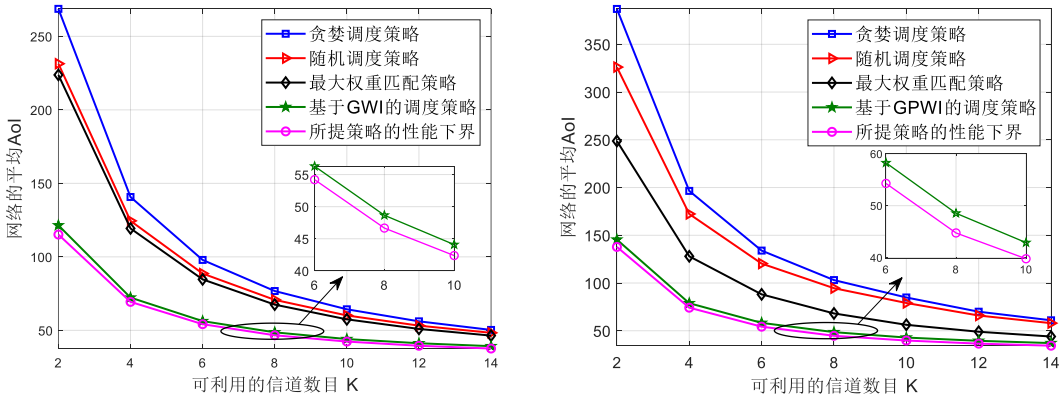
图 4.5 给出了在网络场景一和上述不同残余 AoI 因子下运行 GWI 调度策略，不同信源上的残余 AoI 因子变化曲线。从图 4.5a 和 4.5b 中可以看到，随着时隙数增加，不同信源的残余 AoI 因子 $\alpha_n(t)$ 将趋向于稳定（或周期性变化），这表明 $\mathbb{E}(\alpha_n(t))=\bar{\alpha}_n$。因此，在前文中 $\bar{\alpha}_n$ 代替 $\alpha_n(t)$ 来解耦 CRMAB 问题（4-9）是合理的。为了方便比较，下面仿真中统一使用实时的残余 AoI 因子 $\bar{\alpha}$。



（a）理想信道模型　　（b）非理想信道模型

图 4.6 在图 4.3a 网络场景下所提策略与对比算法的性能随时隙变化的曲线

图 4.6 比较了所提策略和对比算法在网络场景一和不同信道模型下的性能，其中，$N=4$ 和 $K=2$。从图中可以看出，在两种信道模型下，所有调度算法都可以收敛。然而，所提调度策略的性能明显优于其它对比调度策略。此外，所提调度策略

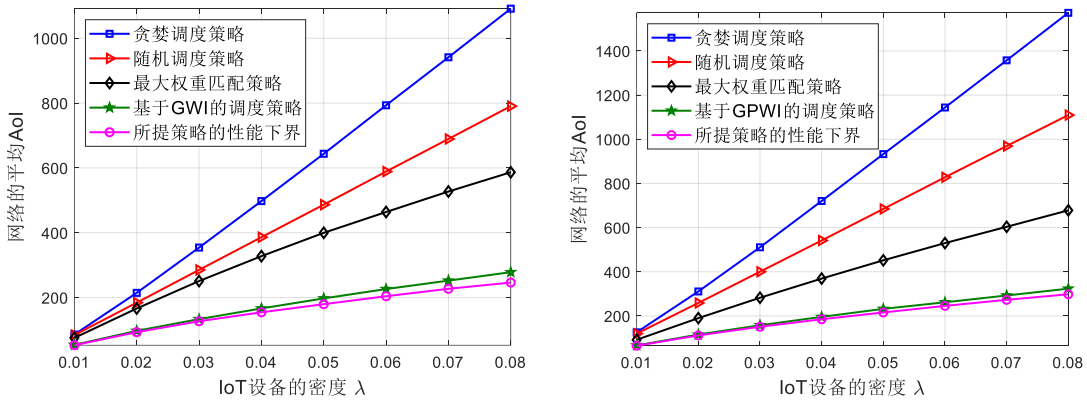


（a）理想信道模型　　（b）非理想信道模型

图 4.7 在图 4.3b 网络场景下所提策略与对比算法的性能随信道数目变化的曲线

的性能接近于性能分析中的理论下界。这些结果表明，基于 GWI 和 GPWI 调度策略的信源不仅考虑了自身的 AoI 和成功传输概率，还考虑与其相关的信源的操作（即参数 $\bar{\alpha}$ ）。

图 4.7 给出了所提策略和对比算法在网络场景二和不同信道模型下的性能曲线，其中，信道数目由 $K=2$ 增加到 $K=14$，以及 $N=30$，$T=500$。从图中可以看到，在两种信道模型下，所有调度策略的网络平均 AoI 都随着信道数目增加而降低。此外，所提策略在对比算法中有最佳性能，并且非常接近的理论的性能下界。然而，当信道数目超过 8 时，网络的性能提升并不显著。这表明，在网络性能和通信资源之间存在一个权衡关系。



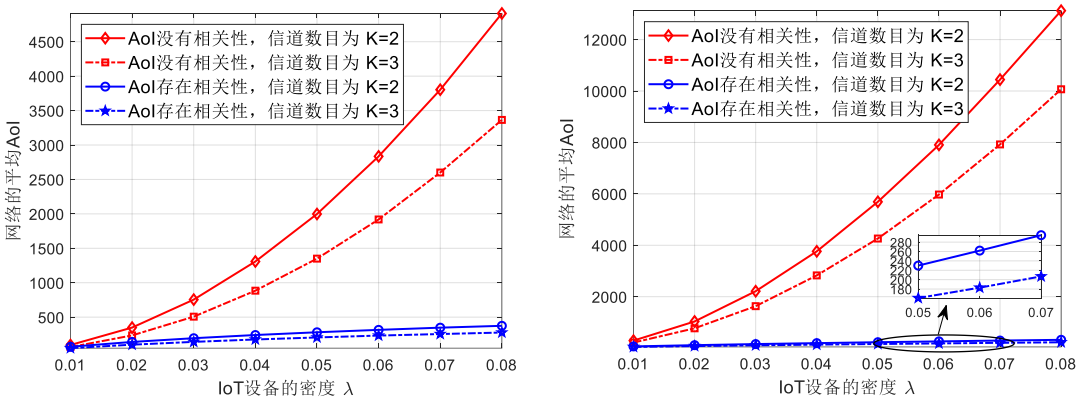
（a）理想信道模型　　（b）非理想信道模型

图 4.8 在动态网络场景下所提策略与对比算法的性能随设备密度变化的曲线

接着，考虑动态的网络场景对算法的影响。在每次蒙特卡洛试验中，传输成功

概率 $\vec{p}$ 和相关矩阵 $\mathcal{E}$ 都根据参数设置中的方式随机生成的。因此，每次蒙特卡洛试验相当于重新产生了一个图 4.3 中的静态网络。图 4.8 比较了不同调度策略在动态网络场景和不同信道模型下网络平均 AoI 随着网络中信源密度 λ 变化的曲线，其中，λ 从 0.01 增加到 0.08，以及 $K=2$，$T=500$。从图中可以看到，在两种信道模型下，所提策略的性能都优于对比调度算法，并且接近理论性能的下界。此外，所提策略与对比调度算法之间的性能差距随着信源密度增加而增加。

最后，考虑信源相关和不相关的网络场景对算法影响。其中，$0<\bar{\alpha}<1$ 对应于信源相关的网络设置；而 $\bar{\alpha}=1$ 对应于信源不相关的设置。图 4.9 给出了在动态网络场景下所提策略的性能随信源密度变化的情况，其中，λ 从 0.01 增加到 0.08，以及 $K=\{2,3\}$，$T=500$。从图中可以看到，在两种信道模型下，所提策略在信源相关网络场景下的性能显著优于不相关网络场景下的性能。此外，这两种网络场景下的性能差距随着信源密度的增加而增加。图 4.9 表明，在两种信道模型下，所提策略通过利用信源之间的相关性来有效地降低网络的平均 AoI。



（a）理想信道模型　　（b）非理想信道模型

图 4.9 在动态网络场景和信源是否相关设定下所提策略与对比算法的性能随设备密度变化的曲线

4.6 本章小结

本章考虑高密度物联网中设备之间存在相关性的情况下基于 AoI 的调度问题。首先，将该网络平均 AoI 最小化问题建模成一个 MDP 问题，但其解存在计算复杂性高、且没有理论性能保证的问题。然后，将该问题建模为一个 CRMAB 问题，并

通过为每一个信源引入待定状态变量将其解耦成 N 个一维的 MAB 问题，从而有效降低问题的计算复杂度。在理想信道和非理想信道下，通过求解每一个 MAB 问题的 Bellman 方程，分别推导了 GWI 和 GPWI 的解析表达式。接着，构造了基于 GWI 和 GPWI 的信源调度策略来解决该 CRMAB 问题。此外，通过求解松弛的拉格朗日问题得到了所提策略的理论性能下界。同时，数值结果验证了该理论分析结果，并表明在高密度网络中所提策略的性能显著优于独立信源下的调度策略。

第 5 章 基于随机 MAB 的水声通信链路自适应机制

本章介绍基于随机 MAB 的水声通信链路自适应机制[①]，研究单条水声通信链路上的传输频率和速率的联合选择问题。首先，通过考虑目标函数和约束条件的特征，提出一种基于模型的随机 MAB 框架。其次，针对平稳水声信道、非平稳水声信道、以及动作空间较大的情况，分别提出三种基于 TS 算法的选择策略求解该基于模型的随机 MAB 问题。最后，理论分析了所提策略的遗憾上界，并通过数值结果验证了所提算法的有效性。

5.1 引言

近年来，水声通信在环境监测、矿产开发与勘探、水下救援、海洋生态系统追踪等领域得到了越来越多的关注[131][132]。然而，与陆地无线通信相比，水声通信面临较低的传播声速（1500 m/s）、有限的带宽、严重的多普勒频移和随机时变的多径传播等挑战。目前，水声领域还没有与陆地无线通信类似的标准信道模型。而且，水下终端（如传感器、自主水下航行器等）通常由电池供电，电池的更换和充电既耗时又昂贵[133]。因此，如何设计高效、可靠的水声通信系统仍是当前一个具有挑战性的问题。

链路自适应技术可以通过自动匹配与当前信道状态信息（Channel State Information，CSI）相一致的传输参数（如调制、编码、发射功率和时频资源）来有效提高链路的传输效率。文献[134]和[135]采用不同的方法来估计或预测水声信道的 CSI，从而自适应地调整链路的信号调制和编码方案。然而，这些工作需要假设水声信道模型已知或服从某种特定的分布，通过估计或预测 CSI 来获得信道与编码方案的一一映射关系。文献[136]考虑远程水声通信中的盲调制与信道编码技术，与文献[134]和[135]不同，作者提出了一种基于信道分类的方法，即通过训练得到信道重要特征，然后根据预测的目标信道来选择最佳的链路调制和编码方案。此外，文献[137]和[138]分别考虑了水下通信中基于功率控制和时频资源分配的链

[①] 本章内容已投稿 IEEE Transaction on Wireless Communications 和部分结果已发表于 IEEE ICC 2020。

路自适应分配策略。

然而，实时的水声 CSI 通常很难获得，使得上述先估计后执行的自适应分配策略不可行。一方面，由于严重的多普勒频移和随机时变的多径传播等特点[139]，水声 CSI 很难通过某些预定或有限的参数来刻画。换言之，由于缺乏统一的水声信道模型，目前还无法建立水声 CSI 与传输参数之间的一一映射关系；另一方面，由于水声传播速度较低，估计或预测的 CSI 并不能反映当前水声信道的真实特性。而且，估计和预测操作通常会导致较高计算复杂度和通信开销，从而降低链路传输效率。

本章研究水声通信链路中的传输频率和速率的联合选择问题，从而最大化链路的平均传输吞吐量。为了克服这些挑战，该问题被建模成随机 MAB 框架，其动作是频率和速率的组合，奖励是传输成功或失败的反馈。在这个序贯决策问题中，本章利用汤普森采样（Thompson Sampling, TS）算法作为选择策略来权衡其中的探索与利用（Exploration and Exploitation，EE）的困境。为了提高传输效率并加快 TS 算法的收敛速度，本章结合目标函数与约束条件的特征，进一步提出了一种基于模型的随机 MAB 框架。其核心思想是通过学习出问题中隐含的模型来序贯地排除明显的次优动作，并持续追踪可能的最优动作。

具体地，本章针对水声通信中的平稳信道、非平稳信道以及大动作空间的情况，分别提出三种相应的选择算法来求解基于模型的 MAB 问题。在平稳信道情况下，利用目标函数的二维单峰特性，提出一种基于单峰目标的汤普森采样（Unimodal Objective TS，UO-TS）算法；在非平稳信道情况下，联合考虑两种非平稳信道模型，即突变信道和缓变信道模型，利用缓变信道追踪和突变信道检测的方法，提出一种混合变化检测的单峰汤普森采样（Hybrid Change Detection UO-TS，HCD-UO-TS）算法；最后，在大动作空间情况下，由于目标函数中的二维单峰特征不存在，通过引入信道容量约束条件，将该联合选择问题进一步描述成带约束的随机 MAB 问题。然后，利用机器学习中的逻辑分类模型，将该约束看成一个分离界面[①]，并通过学习该分离界面来排除明显次优的动作，进而提出了一种迭代边界收缩的汤普森采样（Iterative Boundary Sharking TS, IBS-TS）算法。

[①] 该分离界面将动作空间分成两部分：存在最优动作的集合和不存在最优动作的集合。

5.2 系统模型

考虑单条水声通信链路，其发送端有多个传输频率 $\mathcal{F}=\{f_1, f_2,\ldots, f_m\}$ 和速率 $\mathcal{R}=\{r_1, r_2,\ldots, r_n\}$ 可供选择。假设时间离散化成时隙，即 $t=1,2,\ldots,T$。在每个时隙开始时，发送端选择一对频率和速率进行传输。经过信道后，发送端将接收到来自接收端的反馈（Acknowledgment, ACK）信号（即传输成功或失败的反馈）。系统的目标是通过为链路选择最佳的传输频率和速率来最大化链路的平均吞吐量。

5.2.1 水声信道模型

根据文献[139]，水声信道的特点主要取决于路径损耗增益、环境噪声、大尺度和小尺度效应。其中，路径损耗增益取决于水声传播的物理规律，例如吸收系数和反射系数（底部和表面）；而环境噪声主要由湍流、航运、波浪和热噪声四个来源构成。此外，当海面高度或海底形状发生变化时，传感器的位置会有较大范围的偏移，从而导致水声信道发生变化。具体地，水声信道的路径损耗增益可以看成是与距离 d 和频率 f 相关的函数，即

$$A(d, f)=A_0 d^k a(f)^d \tag{5-1}$$

其中，A_0 是一个单位归一化的常数；k 表示传播因子，用来描述声波传播的几何形状（比如，$k=1$ 和 $k=2$ 分别表示圆柱传播和球形传播，而 $k=1.5$ 表示实际传播）。此外，$a(f)$ 表示吸收系数。利用 Thorp 公式[140]，$a(f)$ 可以表示为（单位：每千米每千赫兹分贝）

$$10\lg a(f)=\frac{0.11 f^2}{1+f^2}+\frac{44 f^2}{4100+f^2}+2.75\times 10^{-4} f^2+0.003 \tag{5-2}$$

其中，$\lg(\cdot)$ 表示以 10 为底的对数。

考虑多径传播、以及传感器位置的大尺度和小尺度的随机位移情况，信道的转移函数可以表示为

$$H(f)=\bar{H}_0(f)\sum_p\sum_i h_{p,i} e^{-j2\pi f\left(\tau_{p,i}-\vartheta_p \tilde{t}\right)} \tag{5-3}$$

其中

$$\bar{H}_0(f)=\frac{1}{\sqrt{A\left(\bar{d}_0,f\right)}} \tag{5-4}$$

和

$$h_{p,i}=\frac{\xi_p}{\sqrt{\left(\frac{d_{p,i}}{\bar{d}_0}\right)^k a(f)^{d_{p,i}-\bar{d}_0}}} \tag{5-5}$$

分别表示距离为 d_0 的参考信道转移函数和第 p 条路径和 i 条散射分量上的信道增益；其中，ξ_p 表示累积反射系数[139]；$d_{p,i}=\bar{d}_p+\Delta_{p,i}$ 表示第 p 条路径的距离，且 $\Delta_{p,i}$ 是由于水面波动和散射导致的位置移动。值得注意的是，该位置变化既包含了小尺度效应 $\delta_{p,i}$（由几个波长引起），又包含了大尺度效应 Δ_p（由许多波长引起），即 $\Delta_{p,i}=\Delta_p+\delta_{p,i}$。此外，$\tau_{p,i}$ 是第 p 条路径上的相对传播延迟，即 $\tau_{p,i}=(d_{p,i}-\bar{d}_0)/c$，其中 $c=1500\,\text{m/s}$ 是声速；$\vartheta_p=v_p/c$ 是对应于速度 v_p 的第 p 条路径上的多普勒因子；$\tilde{t}$ 是在区间 $[0,\ T_{\text{sym}}]$ 内的一个短时间间隔，其中 T_{sym} 表示传输信号的符号周期。

另外，环境噪声的功率谱密度（Power Spectral Density，PSD）可以表示为

$$\begin{cases}10\lg N_t(f)=17-30\lg f\\ 10\lg N_s(f)=40+20(s-0.5)+26\lg f-60\lg(f+0.03)\\ 10\lg N_w(f)=50+7.5w^{1/2}+20\lg f-40\lg(f+0.4)\\ 10\lg N_{th}(f)=-15+20\lg f\end{cases} \tag{5-6}$$

其中，w 表示风速（单位为：m/s）；s 表示船舶活动因子，通常在区间 $[0,\ 1]$ 之间变化。因此，总的噪声功率谱密度为 $N(f)=N_t(f)+N_s(f)+N_w(f)+N_{th}(f)$。

5.2.2 最优化问题模型

信号经过水声信道传播后，接收端的瞬时信噪比（Signal-to-Noise Ratio, SNR）可以表示为

$$\Psi(f_i)=\int_{f_i}^{f_i+B}\frac{P\,|H(f)|^2}{N(f)B}df \tag{5-7}$$

其中，P 表示发送功率；B 表示信号带宽（单位为：赫兹）。然而，在实际的水声通信系统中，一条链路通常只能支撑有限个数的传输速率。根据文献[82]，每个传

输速率r_j将对应一个传输成功概率θ_{f_i,r_j}，且该传输成功概率由瞬时 SNR 决定，即

$$\theta_{f_i,r_j} \triangleq \Pr\{\Psi(f_i) \geq \Psi_{r_j}^{\text{ref}}\} \tag{5-8}$$

其中，$\Psi_{r_j}^{\text{ref}}$表示链路采用频率f_i和速率r_j进行传输时，接收端成功解调该信号需要的最小接收 SNR。因此，链路的平均吞吐量可以表示为$r_j \times \theta_{f_i,r_j}$。

最后，该传输频率和速率的联合选择问题可以建模成

$$\begin{aligned} \max_{f_i,r_j} \quad & r_j \times \theta_{f_i,r_j} \\ \text{s.t.} \quad & f_i, r_j \in \mathcal{F} \otimes \mathcal{R} \end{aligned} \tag{5-9}$$

其中，$\otimes$表示笛卡尔集合运算符。从上式可以看出，若θ_{f_i,r_j}的解析表达式已知，问题（5-9）可以采用传统的优化方法进行求解。然后，由于目前没有一个统一的水声信道来对复杂的海洋环境进行建模，所以θ_{f_i,r_j}的解析表达式无法准确获取。为了克服该挑战，本文利用在线学习方法为水声链路序贯地学习出最佳的传输频率和速率。在这个过程中，发送端需要仔细权衡 EE 困境：一方面，链路需要传输在当前最佳的频率和速率对上，以期最大化其累积奖励；另一方面，它需要探索新的频率和速率对，以免错过最优解，从而最大化其长期奖励。因此，下面将问题（5-9）建模成一个随机 MAB 问题。

5.2.3 随机 MAB 问题模型

（1）随机 MAB 的相关概念与术语

在该随机 MAB 问题中，玩家是水声链路的发送端；动作是频率和速率的组合，表示为$\mathcal{A} = \{a_{1,1}, a_{1,2}, \ldots, a_{m,n}\}$；奖励是传输成功与失败的反馈。在每个时隙$t$，玩家首先从集合$\mathcal{A}$中选择一个动作$I_a^t$进行数据传输，之后，它会观察到一个反馈或奖励$X_{I_a^t}(t) \in \{0,1\}$；当$X_{I_a^t}(t) = 1$时，表示该次数据传输成功；否则，$X_{I_a^t}(t) = 0$。多个回合之后，动作$a_{i,j}$上估计的均值可以表示为$\theta_{a_{i,j}} = \mathbb{E}[X_{a_{i,j}}(t)]$，其中，$\mathbb{E}[\cdot]$表示取期望运算。因此，问题（5-9）可以从重新表示成

$$\begin{aligned} a^* &= \arg\max_{f_i, r_j} \quad r_j \times \theta_{f_i, r_j} \\ &= \arg\max_{f_i, r_j} \quad r_j \times \mathbb{E}[X_{a_{i,j}}(t)] \end{aligned} \tag{5-10}$$

其中，a^* 表示集合 $\mathcal{A}$ 中的最优动作。

下面介绍伪遗憾这一概念来衡量所提算法的性能。令 $\mu_{a_{i,j}} = r_j \times \theta_{a_{i,j}}$ 表示动作 $a_{i,j}$ 上估计的平均吞吐量，则伪遗憾可以表示成

$$\begin{aligned} \mathcal{R}eg_T &= \max_{a_{i,j} \in \mathcal{A}} \mathbb{E}\left[r_j \sum_{t=1}^{T} X_{a_{i,j}}(t) - r_j \sum_{t=1}^{T} X_{I_a^t}(t) \right] \\ &= T \times r^* \theta^* - r_j \mathbb{E}\left[\sum_{t=1}^{T} \theta_{I_a^t} \right] \\ &= T\mu^* - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{I_a^t} \right] \end{aligned} \tag{5-11}$$

其中，θ^* 和 μ^* 分别表示最佳动作的成功传输概率和平均吞吐量；上式中第一个期望运算作用于随机变量 $X_{I_a^t}(t)$，而第二个和第三个期望运算都作用于随机选择策略上。令 $D_{a_{i,j}}(t)$ 表示动作 $a_{i,j}$ 到时间 t 为止被玩家选中的次数，则上式可以简化为

$$\mathcal{R}eg_T = \sum_{a_{i,j} \in \mathcal{A}} \Delta_{a_{i,j}} \mathbb{E}\left[D_{a_{i,j}}(T) \right] \tag{5-12}$$

其中，$\Delta_{a_{i,j}} = \mu^* - \mu_{a_{i,j}}$。

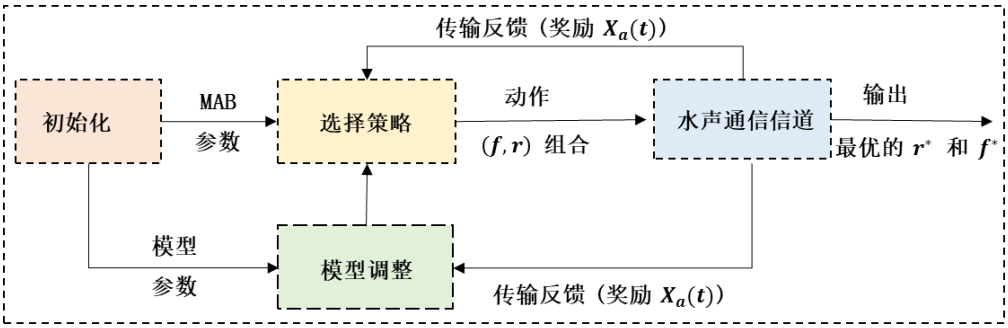


图 5.1 基于模型的随机 MAB 框架

（2）基于模型的随机 MAB 框架

但动作空间较大时，传统的随机 MAB 问题面临收敛速度慢的问题。因此，本节结合问题模型的特征，进一步提出一种基于模型的随机 MAB 框架，如图 5.1 所

示。与传统的随机 MAB 框架相比，所提框架中的奖励不仅用于评估下一轮的选择策略，还可以用来学习问题（5-9）中目标函数和约束条件的一些特征。从图 5.1 中可以看到，基于模型的随机 MAB 框架主要包含四部分：初始化、选择策略、水声通信信道以及模型调整。下面对其进行分别阐述：

- 初始化：这部分的作用主要包括初始化选择策略和模型调整模块中的参数。值得注意的是，并不是所有的参数都需要初始化为随机数或零，而是可以设置为从实际环境中获取的一些经验值。例如，在模型调整部分，某个海域的扩频因子 k 、风速 w 和航运活动因子 s 都可以采用实测值作为初始化参数，从而加快算法的收敛速度。
- 选择策略：这部分是所提框架的核心，它通过采用某种动作选择策略来有效地权衡 EE 困境。在绪论中介绍到，对于随机 MAB 问题，其求解算法主要包括 ε-贪婪算法、UCB 算法、KL-UCB 算法和 TS 算法。在该链路自适应问题中，由于奖励过程可以看出是伯努利分布，传输成功概率则可以表示成 Beta 分布。因此，可以用 TS 算法作为选择策略来为链路寻找最优的传输频率和速率组合。
- 水声通信信道：为了模拟真实的水声通信环境，比如多径、强表面散射、多普勒频移和随机位置位移等特征，这部分采用文献[139]中的信道模型对所提选择策略进行评估。
- 模型调整：这部分的作用是学习出关于问题（5-9）中目标函数和约束条件的一些特征信息。例如，此信息可以是目标函数的结构、水声信道模型的特征以及数据速率和频率之间的关系。然后，将学习到的特征信息反馈到选择策略中，指导下一轮的选择。通过这种方式，玩家不仅可以排除一些明显次优的动作，又可以快速找到空间中的最优动作。

5.3 基于随机 MAB 的联合选择策略

本节考虑水声信道的平稳和非平稳特征，以及动作空间较大的情况，分别提出 UO-TS、HCD-UO-TS 和 IBS-TS 算法来为链路寻找最佳的传输频率和速率，从而提高水声链路的传输效率。

5.3.1 平稳信道下的 UO-TS 选择策略

平稳的水声信道通常存在于相对静止的环境中，如水池和小型湖泊等。在这些场景下，通常认为信道的增益是平稳的。本节首先给出平稳的速率-信道模型和频率-信道模型；其次，分析问题（5-9）中目标函数的二维单峰特性；最后，基于该特性提出 UO-TS 算法来求解上一节基于模型的随机 MAB 问题。

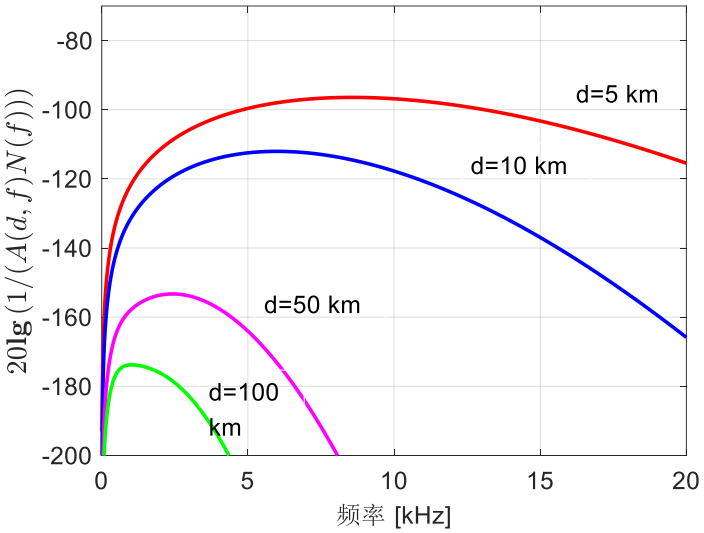


图 5.2 水声信道增益随频率变化的曲线

（1）频率相关性的信道模型

在平稳环境中，接收信号的平均信噪比可以看成保持不变，即不随时间变化而变化。此时，随机位置位移和多普勒频移的过程可以看作是平稳过程，且信道的增益主要取决于距离和频率相关的衰减。所以，式（5-3）中的信道传递函数可以简写为

$$H(f)=\bar{H}_0(f)\sum_p \bar{h}_p e^{-j2\pi f\bar{\tau}_p} \tag{5-13}$$

其中，$\bar{h}_p$ 和 $\bar{\tau}_p$ 分别表示第 p 条路径上的平均信道增益和时延。根据式（5-7），接收信噪比可以看成是关于频率 f 的函数。由文献[139]可知，$a(f)^{d_p-\bar{d}_0}\approx a^{d_p-\bar{d}_0}$，因此接收信噪比取决于表达式$\left(A(\bar{d}_0,f)N(f)\right)^{-1}$。根据香农公式，传输误码率（Bit Error Rate, BER）与平均信噪比成反比，即与$\left(A(\bar{d}_0,f)N(f)\right)$成正比。图 5.2 给出了表达

式$\left(A(\bar{d}_0, f)N(f)\right)^{-1}$在距离$d=\{5, 10, 50, 100\}$ km 的频率特性曲线。从图中可以看到，当给定传输距离和速率时，存在唯一的频率使得接收信噪比最大化。因此，传输成功概率与频率存在如下关系

$$\theta_{f_i} \propto \frac{1}{\bar{A}(d, f_i)N(f_i)}, \quad i=1,2,\ldots,m \tag{5-14}$$

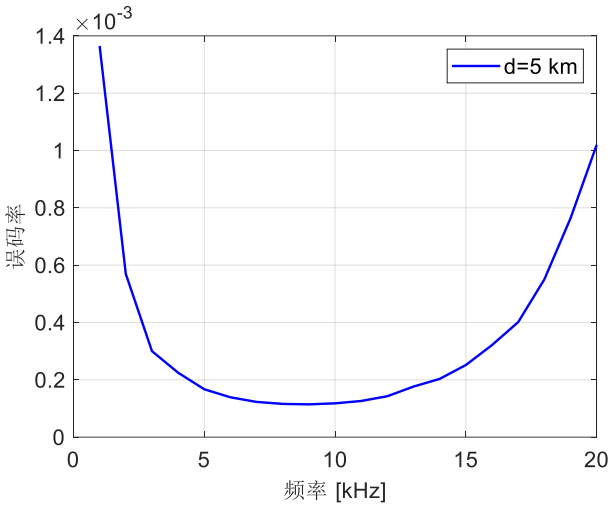


图 5.3 误码率随着频率变化的曲线

为了验证上述关系，基于文献[139]中的信道模型，图 5.3 给出了误码率随频率变化的曲线。其中，信号采用 QPSK 与 OFDM 的调制方式，传输距离为 5 千米，水声信道参数为$k=1.5$，$s=0.5$，$w=0$。从图中可以看到，误码率是关于频率的一个近似的拟凸函数；这表明在给定传输速率下，存在唯一的频率f^*使得传输成功概率最大化。

（2）速率相关性的信道模型

假设信道有n个可能的状态，并按升序排列为$h_1<h_2<\ldots<h_n$；在每个时隙，信道处于状态i的概率为P_{h_i}，且$\sum_i P_{h_i}=1$。值得注意的是，状态h_i的分布未知。根据香农公式，每个信道状态h_i对应一个最大传输速率r_i。因此，对于传输速率r_i，其对应的传输成功概率为

$$\theta_{r_i}=\sum_{j=i}^{n} P_{h_j} \tag{5-15}$$

换言之，传输速率越高其传输成功概率将越低，即两者成反比关系：

$$\theta_{r_j} \propto \frac{1}{r_j}, \quad j=1,2,\ldots,n \tag{5-16}$$

其中，$\theta_{r_1} > \theta_{r_2} > \ldots > \theta_{r_n}$。

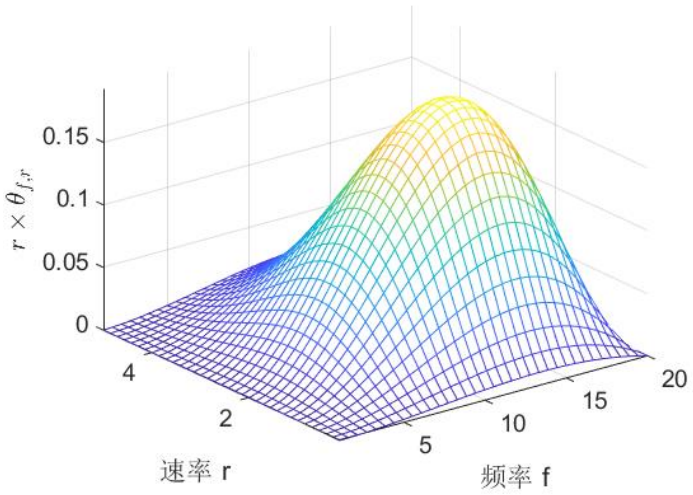


图 5.4 目标函数的二维单峰图像

下面讨论问题（5-9）中目标函数的二维单峰特性。首先，利用无向图$\mathcal{G}=(\mathcal{V},\mathcal{E})$来描述单峰特性，其中，$\mathcal{V}$和$\mathcal{E}$分别表示图的顶点和边的集合。在该无向图中，每个顶点代表随机 MAB 问题中的一个动作；若一个动作可以转移到另一个动作是，则它们之间存在一条相连的边。在本文中，无向图中的每一顶点仅与其邻居节点相连。根据文献[78]，在给定最优动作的情况下，无论起始顶点为无向图$\mathcal{G}$中的某处，都存在一条路径使其到达最优动作所在的顶点。

结合式（5-14）和（5-16），问题（5-9）可以重新表示成

$$\begin{aligned} \max_{f_i, r_j} \quad & r_j \times \theta_{f_i} \times \theta_{r_j} \\ \text{s.t.} \quad & f_i \in \mathcal{F} \text{ and } r_j \in \mathcal{R} \end{aligned} \tag{5-17}$$

一方面，当数据速率固定时，成功传输概率是频率f的拟凹函数（即单峰结构）。这一点可以通过公式（5-14）和图 5.3 中的频率相关特征得到验证；另一方面，当传输频率给定时，成功传输概率与速率呈线性变化，即公式（5-15）。因此，其平均吞吐量$r\theta_r$具有单峰结构。为了验证这一点，图 5.4 给出了问题（5-9）中目标函

数的图像，仿真中的参数根据公式（5-14）与（5-15）设定。从图中可以看到，问题（5-9）的目标函数是一个拟凹函数，即具有二维单峰结构。

算法 5.1 基于 UO-TS 的选择策略

步骤 1： 初始化参数： $m,n,\mathcal{R},\mathcal{F},T$ ； 同时设定 $\hat{\mu}_a(0)=0, D_a^0=0, \alpha_a=0,$ $\beta_a=0, \forall a\in\mathcal{A}$

步骤 2: 执行 $t=1,2,...,T$ 次以下步骤

步骤 3: 找出动作空间中的 leader： $L(t)=\arg\max_{a\in\mathcal{A}}\hat{\mu}_a(t)$

步骤 4: 利用 Beta 分布获取集合 $\mathcal{M}^+(L(t))$ 中所有动作的经验成功概率： $\tilde{\theta}_a \sim \text{Beta}(\alpha_a+1,\beta_a+1), \forall a\in\mathcal{M}^+(L(t))$

步骤 5: 确定当前的动作： $I_a^t=\arg\max_{a\in\mathcal{M}^+(L(t))} r_a\tilde{\theta}_a$

步骤 6： 利用动作 I_a^t 上的频率和速率进行传输，并观测传输反馈 $X_{I_a^t}(t)$

步骤 7： 利用公式（5-18）更新动作 I_a^t 的经验平均值 $\hat{\mu}_{I_a^t}(t)$

步骤 8： 更新动作 I_a^t 的执行次数： $D_{I_a^t}(t)=D_{I_a^t}(t)+1$

步骤 9： 更新动作 I_a^t 的 Beta 分布参数：当 $X_{I_a^t}(t)=1$ 时， $\alpha_{I_a^t}=\alpha_{I_a^t}+1$ ；否则， $\beta_{I_a^t}=\beta_{I_a^t}+1$

（3）所提的 UO-TS 算法

基于该二维单峰结构，本节提出一种 UO-TS 算法来求解随机 MAB 问题，如算法 5.1 所示。令 $\mathcal{M}(a)=\{a'\in\mathcal{A} \text{ and } (a,a')\in\mathcal{E}\}$ 表示动作 a 的相邻动作集合；同时，集合 $\mathcal{M}^+(a)=\{\mathcal{M}(a)\}\cup\{a\}$ 表示关于动作 a 的所有动作集合。根据 TS 算法，假设每个动作上的奖励服从贝努利分布，而传输成功概率服从 Beta 分布。考虑到开始时算法需要进行均匀探索，因此 Beta 分布的参数初始化为 $\alpha=0$ 和 $\beta=0$ 。在算法 5.1 中，首先根据每个动作上的经验平均奖励寻找当前时刻的 leader，即 $L(t)=\arg\max_{a\in\mathcal{A}}\hat{\mu}_a(t)$ ，其中，动作 a 在时间 t 的经验平均奖励可以表示为

$$\hat{\mu}_a(t)=\hat{\mu}_a(t-1)+\frac{1}{D_a(t-1)}\left(r_a X_a(t-1)-\hat{\mu}_a(t-1)\right) \tag{5-18}$$

其中，r_a 和 $D_a(t)$ 分别表示动作 a 的传输速率和直至时间 t 被选中的次数。确定 leader 后，算法从 Beta 分布中产生集合 $\mathcal{M}^+(L(t))$ 中动作的经验成功概率 $\tilde{\theta}_a$；接着，根据集合 $\mathcal{M}^+(L(t))$ 中动作的平均吞吐量，选择当前时隙要执行的动作 I_a^t，即 $I_a^t=\arg\max_{a\in\mathcal{M}^+(L(t))} r_a\tilde{\theta}_a$；然后，链路利用动作 I_a^t 上的速率和频率进行数据传输，并得到传输成功或失败的反馈，作为当前时隙的奖励 $X_{I_a^t}(t)$。最后，根据接收到的奖励，算法第 7 步到第 9 步分别对动作 I_a^t 上的平均奖励 $\hat{\mu}_{I_a^t}(t)$、被选择次数 $D_{I_a^t}(t)$ 以及其 Beta 分布参数 (α,β) 进行更新。

5.3.2 非平稳信道下的 HCD-UO-TS 选择策略

本节考虑两种非平稳信道，即突变信道和缓变信道。针对突变信道，采用广义似然率（GLR）检验进行断点检测；针对缓变信道，提出一种新颖的加权平均方法（RERWA）。最后，通过联合这两个方法来对抗非平稳信道，提出一种 HCD-UO-TS 算法来求解上述随机 MAB 问题。

（1）突变信道中基于 GLR 的断点检测方法

突变信道也被称为块衰落或准静态信道，它假设信道的平均增益在时间块内保持恒定，但在块间发生快速变化。在水声通信中，这种情况通常出现在海平面高度或底部形状突然变化时，且需要经历较长的时间才会出现一个突变点。由于成功传输概率与信道增益成正比，因此可以通过判断传输成功概率是否存在突变点来检测信道是否发生突变。在文献中，突变点检测方法主要有 CUSUM 检验、KS 检验以及 GLR 检验。其中，CUSUM 检验是一种参数检验，它需要已知样本的先验分布和参数数值。但是这种先验信息在水声通信中很难获得，因为唯一观察到的信息是一个二元传输反馈。虽然观测的奖励可以看作是伯努利分布，但分布的参数值确是未知的。此外，KS 检验是一种非参数检验，不需要样本的任何先验信息，但是它的性能相对较差，且在样本量很小的情况下无法有效检测出突变点。GLR 检验介于两者之间，只需要知道样本的分布，而不需要知道分布参数的值；且少量样

本就有较好的性能。因此，本节采用 GLR 检验来检测水声信道中的突变点。

具体地，令 $X(t)$ 表示动作直至时间 t 获得样本序列。根据文献[142]和贝努利分布奖励，GLR 的检验统计量可以构造为

$$\mathcal{T}_{\mathrm{GLR}}(t)=\sup_{n\in[1,t]}\left[n\times\mathrm{KL}(\hat{\theta}_{1:n},\hat{\theta}_{1:t})+(t-n)\times\mathrm{KL}(\hat{\theta}_{n+1:t},\hat{\theta}_{1:t})\right] \tag{5-19}$$

其中，$\mathrm{KL}(\cdot)$ 表示 Kullback-Leibler 散度；同时，$\hat{\theta}_{1:t}=\sum_{i=1}^{t}X(i)\big/t$ 表示前 t 个历史样本的平均传输概率。接着，根据奈曼-皮尔逊准则，在给定虚警概率 P_f 的情况下，GLR 检验统计量的门限表达式为

$$\mathcal{T}_{\mathrm{ref}}=\ln\left(\frac{3t\sqrt{t}}{P_f}\right) \tag{5-20}$$

因此，当检验统计量大于门限时，即 $\mathcal{T}_{\mathrm{GLR}}\geq\mathcal{T}_{\mathrm{ref}}$，便认为样本中存在突变点；否则，样本中不存在突变点。

（2）缓变信道中基于 RERWA 的非平稳追踪方法

缓变信道通常指信道的平均增益随着时间连续缓慢地变化，且这种变化不能被上述 GLR 检验方法所检测。在文献中，有两种处理缓变信道的方法：加窗平均法（Sliding Window, SW）和新近指数加权平均法（Exponential Recency Weighted Average, ERWA）。SW 方法的核心思想是只对时间窗内的样本求平均值，即

$$\hat{\mu}_a^{\mathrm{SW}}(t+1)=\sum_{i=t-\varsigma}^{t}\frac{r_aX_a(i)}{\mathbf{1}\{I_a^i=a\}} \tag{5-21}$$

其中，ς 表示时间窗的长度。另外，ERWA 方法的核心是给靠近当前时刻的样本大的权重，远离当前时刻的样本小的权重，即

$$\begin{aligned}\hat{\mu}_a^{\mathrm{EW}}(t+1)&=\hat{\mu}_a^{\mathrm{EW}}(t)+\gamma\left(r_aX_a(t)-\hat{\mu}_a^{\mathrm{EW}}(t)\right)\\&=(1-\gamma)^t\hat{\mu}_a^{\mathrm{EW}}(0)+\sum_{i=1}^{t}\gamma(1-\gamma)^{t-i}r_aX_a(i)\end{aligned} \tag{5-22}$$

其中，γ 是加权因子。从上式可以看到，ERWA 方法是关于样本均值的一种有偏估计，其样本平均依赖于初始值 $\hat{\mu}_a^{\mathrm{EW}}(0)$。为了克服这种情况，通过引入变化的加权因子，本文提出一种改进的 ERWA（Refined ERWA, RERWA）方法来对抗缓变的

水声信道，即

$$\hat{\mu}_a^{\mathrm{REW}}(t+1)=\hat{\mu}_a^{\mathrm{REW}}(t)+\frac{\dot{\gamma}}{\bar{\iota}(t)}\left(r_a X_a(t)-\hat{\mu}_a^{\mathrm{REW}}(t)\right) \tag{5-23}$$

其中，$\bar{\iota}(t)=\bar{\iota}(t-1)+\dot{\gamma}\left(1-\bar{\iota}(t-1)\right)$ 且 $\iota(0)=0$。

性质 5.1. 所提 ERWA 方法是关于样本平均 $\hat{\mu}_a(t)$ 的无偏估计，即式（5-23）不依赖于初始值 $\hat{\mu}_a(0)$。

证明：上述性质等价于证明式（5-23）中的权重之和等于 1。为了方便，令 $S_t=\dot{\gamma}/\bar{\iota}(t)$ 表示 RERWA 方法在时隙 t 的变化加权因子，则式（5-23）可以重新表示为

$$\begin{aligned}\hat{\mu}_a^{\mathrm{REW}}(t+1)&=\hat{\mu}_a^{\mathrm{REW}}(t)+S_t\left(r_a X_a(t)-\hat{\mu}_a^{\mathrm{REW}}(t)\right)\\&=S_t r_a X_a(t)+\left(1-S_t\right)\hat{\mu}_a^{\mathrm{REW}}(t)\end{aligned} \tag{5-24}$$

将项 $\hat{\mu}_a^{\mathrm{REW}}(t)$ 从 0 到 t 时刻展开，可以得到

$$\begin{aligned}\hat{\mu}_a^{\mathrm{REW}}(t+1)&=r_a S_t X_a(t)+(1-S_t)r_a S_{t-1}X_a(t-1)+\cdots\\&\quad+(1-S_t)(1-S_{t-1})\cdots(1-S_1)\hat{\mu}_a^{\mathrm{REW}}(1)\\&=S_t r_a X_a(t)+r_a\sum_{i=1}^{t-1}S_i X_a(i)\prod_{k=i+1}^{t}(1-S_k)+\prod_{j=1}^{t}(1-S_j)\hat{\mu}_a^{\mathrm{REW}}(1)\end{aligned} \tag{5-25}$$

因此，下面只需证明

$$\underbrace{S_t}_{Z_1}+\underbrace{\sum_{i=1}^{t-1}S_i\prod_{k=i+1}^{t}(1-S_k)}_{Z_2}+\underbrace{\prod_{j=1}^{t}(1-S_j)}_{Z_3}=1 \tag{5-26}$$

在证明上式前，需要用到以下关系。首先，将项 $\bar{\iota}(t)$ 从 0 到 t 时刻展开，得到

$$\begin{aligned}\bar{\iota}(t)&=(1-\dot{\gamma})^t\bar{\iota}(0)+\sum_{i=1}^{t}\dot{\gamma}(1-\dot{\gamma})^{t-i}\\&=\sum_{i=1}^{t}\dot{\gamma}(1-\dot{\gamma})^{t-i}\\&=1-(1-\dot{\gamma})^t\end{aligned} \tag{5-27}$$

其中，第一个等号是因为 $\bar{\iota}(0)=0$；第二个等号是利用了等数列求和公式。接着，可以得到

$$S_t = \frac{\dot{\gamma}}{1-(1-\dot{\gamma})^t} \tag{5-28}$$

和

$$1-S_t = (1-\dot{\gamma})\frac{1-(1-\dot{\gamma})^{t-1}}{1-(1-\dot{\gamma})^t} \tag{5-29}$$

因此，对于式（5-26）的第三项 Z_3，可以得到以下关系：

$$Z_3 = (1-\dot{\gamma})\frac{1-(1-\dot{\gamma})^0}{1-(1-\dot{\gamma})^1}\times(1-\dot{\gamma})\frac{1-(1-\dot{\gamma})^1}{1-(1-\dot{\gamma})^2}\times\cdots\times(1-\dot{\gamma})\frac{1-(1-\dot{\gamma})^{k-1}}{1-(1-\dot{\gamma})^k} = 0 \tag{5-30}$$

这表明所提的 RERWA 方法不依赖于初始值。对于第二项 Z_2，利用式（5-28）和（5-29），可以得到

$$\begin{aligned} Z_2 &= \sum_{i=1}^{t-1} S_i \times (1-\dot{\gamma})\frac{1-(1-\dot{\gamma})^i}{1-(1-\dot{\gamma})^{i+1}}\times(1-\dot{\gamma})\frac{1-(1-\dot{\gamma})^{i+1}}{1-(1-\dot{\gamma})^{i+2}}\times\cdots\times(1-\dot{\gamma})\frac{1-(1-\dot{\gamma})^{t-1}}{1-(1-\dot{\gamma})^t} \\ &= \sum_{i=1}^{t-1}\frac{\dot{\gamma}}{1-(1-\dot{\gamma})^i}\times(1-\dot{\gamma})^{t-i}\times\frac{1-(1-\dot{\gamma})^i}{1-(1-\dot{\gamma})^t} \\ &= 1-\frac{\dot{\gamma}}{1-(1-\dot{\gamma})^t} \end{aligned} \tag{5-31}$$

另外，

$$Z_1 = \frac{\dot{\gamma}}{1-(1-\dot{\gamma})^t} \tag{5-32}$$

因此，结合公式（5-30）、（5-31）和（5-32），可以得到 $Z_1+Z_2+Z_3=1$。 □

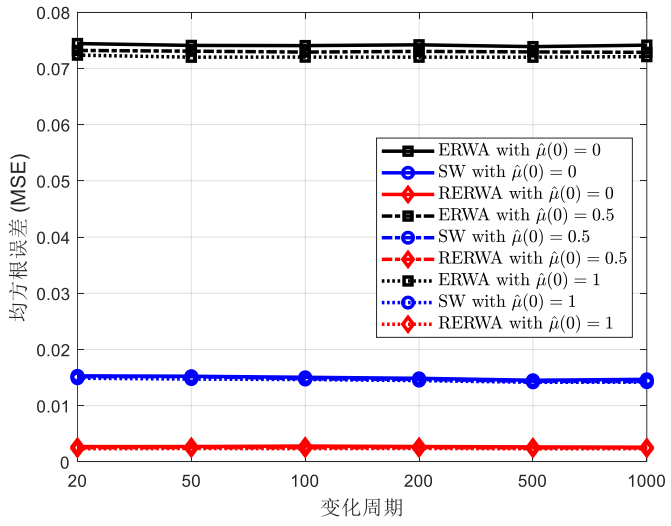


图 5.5 不同非平稳检测方法在不同条件下的均方根误差曲线

为了验证 RERWA 方法的有效性，在考虑不同初始值和变化周期的情况下，图 5.5 对比了三种方法的均方误差随变化周期改变的曲线。所有结果由 1000 次蒙特卡洛仿真得到，其中非平稳信道参数设置为：每个动作上的平均值只在变化周期之间发生缓慢变化，且变化范围在$[\mu_a-0.1,\mu_a+0.1]$之间。此外，SW 方法的时间窗设定为$\varsigma=100$；ERWA 方法和 RERWA 方法的加权因子分布设定为$\gamma=0.1$和$\dot{\gamma}=10^{-4}$。从图中可以看到，所提 RERWA 方法拥有最佳性能，且不受初始值影响。

算法 5.2 基于 HCD-UO-TS 的选择策略

步骤 1： 初始化参数： $m,n,\varsigma,\gamma,\nu,\mathcal{T}_{\text{ref}},\mathcal{R},\mathcal{F},T$，同时设定 $\hat{\mu}_a(0)=0,D_a^0=0,$ $\alpha_a=0,\beta_a=0,\forall a\in\mathcal{A}$

步骤 2: 执行$t=1,2,...,T$次以下步骤

TS 算法部分：

步骤 3: 找出动作空间中的 leader： $L(t)=\arg\max_{a\in\mathcal{A}}\hat{\mu}_a(t)$

步骤 4: 利用 Beta 分布获取集合$\mathcal{M}^+(L(t))$中所有动作的经验成功概率：

$\tilde{\theta}_a\sim\text{Beta}(\alpha_a+1,\beta_a+1),\forall a\in\mathcal{M}^+(L(t))$

步骤 5: 确定当前的动作： $I_a^t=\arg\max_{a\in\mathcal{M}^+(L(t))}r_a\tilde{\theta}_a$

步骤 6： 利用动作I_a^t上的频率和速率进行传输，并观测传输反馈$X_{I_a^t}(t)$

步骤 7： 更新动作I_a^t的执行次数： $D_{I_a^t}(t)=D_{I_a^t}(t)+1$

步骤 8： 更新 Beta 参数：当$X_{I_a^t}(t)=1$时， $\alpha_{I_a^t}=\alpha_{I_a^t}+1$；否则， $\beta_{I_a^t}=\beta_{I_a^t}+1$

缓变信道跟踪部分：

步骤 9： 利用公式（5-23）来更新动作I_a^t的经验平均值$\hat{\mu}_{I_a^t}(t)$

突变信道检测部分：

步骤 10： 如果$D_a(t)>\nu$，则继续以下步骤；否则，跳转到步骤 3

步骤 11： 利用公式（5-20）来计算 GLR 检测法的检验统计量$\mathcal{T}_{\text{GLR}}$

步骤 12： 如果$\mathcal{T}_{\text{GLR}}\geq\mathcal{T}_{\text{ref}}$，则继续以下步骤；否则，跳转到步骤 3

步骤 13： $D_a(t)=0,\alpha_a=0$和$\beta_a=0,\forall a\in\mathcal{A}$

（3）所提出的 HCD-UO-TS 算法

最后，基于 GLR 检验方法和 RERWA 方法，本节提出一种 HCD-UO-TS 算法来求解随机 MAB 问题，如算法 5.2 所示[①]。从算法 5.2 可以看到，HCD-UO-TS 算法主要包含三个部分：TS 算法、缓变信道跟踪和突变信道检测。其中，TS 算法部分与 UO-TS 算法相同。对于缓变信道追踪，采用公式（5-23）中的 RERWA 方法来更新动作的经验平均吞吐量。对于突变信道检测，首先设计了一个长度为ν的时间窗，当某个动作上的样本长度超过该时间窗时，算法便调用 GLR 检验法来检测当前样本中是否存在突变点。通过与设定好的门限$\mathcal{T}_{ref}$比较，若 GLR 检验统计量超过该门限，则认为样本中存在突变点，需要重新初始化 TS 算法中的参数；否则，算法继续运行。

与 UO-TS 算法相比，HCD-UO-TS 算法主要有两个特点。第一，HCD-UO-TS 算法利用 RERWA 方法来更新动作的经验平均吞吐量，而不是传统的样本平均值方法；第二，HCD-UO-TS 算法一旦检测到突变点，便立刻重新初始化 TS 算法中所有动作上的参数。因此，HCD-UO-TS 算法更能适应真实水声通信环境。

5.3.3 大动作空间下的 IBS-TS 选择策略

在 5.3.1 和 5.3.2 节中，本文假设问题的目标函数具有二维单峰特性。但是，当动作空间较大时，将出现很多动作的成功传输概率为 0 或为 1 的情况，导致目标函数的单峰特性消失。针对这种情况，本节结合信道容量公式，将问题（5-9）重新建模成一个带约束的最优化问题，即

$$\begin{aligned} \max_{f_i, r_j} \quad & r_j \times \theta_{f_i, r_j} \\ \text{s.t.} \quad & r_j \le B \log_2 \left(1 + \Psi(f_i)\right) \\ & f_i, r_j \in \mathcal{F} \otimes \mathcal{R} \end{aligned} \tag{5-33}$$

其中，$\Psi(f_i)$是接收信噪比。类似线性规划的思想，第一个约束条件可以看作是一个分离界面将动作空间分成两部分：存在最优动作的集合和不存在最优动作的集

[①] 需要注意的是，本节虽然考虑非平稳信道，但仍假设问题中目标函数的单峰特性成立，且其最佳动作可能随着信道突变而发生变化。每当算法检测到突变点时，立刻重新初始 TS 算法中所有动作上的参数。

合。若 TS 算法只在其中一个动作集合中进行探索，便可以很大程度上加快算法的收敛速度。为了实现这一目标，本节基于逻辑回归的思想，提出一种 IBS-TS 算法来解决问题（5-33）。在逻辑回归模型中，分离界面通常是一个线性函数。由于问题（5-33）中的约束条件是一个非线性函数，因此需要对其进行线性转换。

性质 5.2. 令 $c_5=-\lg\left(P\left|\sum_p\sum_i h_{p,i}\right|^2\Big/A_0\bar{d}_0^{k}B10^{N_1}\right)$，$c_2=8\pi^2\bar{\tau}^2/\ln 10, c_3=-\eta$ 以及 $c_1=c_4=1$，其中 $\bar{\tau}$ 表示传播路径的平均时延，则存在一个线性分离平面将动作空间分成两部分，且该线性平面表达式为

$$c_1\varphi_1(r)+c_2\varphi_2(f)+c_3\varphi_3(f)+c_4\varphi_4(f)+c_5=0 \tag{5-34}$$

其中，$\varphi_1(r)=\lg\left(2^{r/B}-1\right), \varphi_2(f)=f^2, \varphi_3(f)=\lg f, \varphi_4(f)=\lg a(f)^{\bar{d}_0}$。

证明：不失一般性，考虑窄带信号，其接收信噪比表达式为

$$\Psi(f)=\frac{P|H(f)|^2}{N(f)B} \tag{5-35}$$

利用香农公式，可以得到传输速率与频率之间的关系

$$r=B\log_2\left(1+\Psi(f)\right) \tag{5-36}$$

经简单变换，可得

$$2^{r/B}-1=\frac{|H(f)|^2\,P}{N(f)B} \tag{5-37}$$

两边取对数得到

$$\lg\left(2^{r/B}-1\right)=\lg|H(f)|^2-\lg N(f)+\lg\frac{P}{B} \tag{5-38}$$

对于上式右手第一项，可知

$$\begin{aligned}|H(f)|^2&=\left|\bar{H}_0(f)\sum_p\sum_i h_{p,i}e^{-j2\pi f\hat{\tau}_{p,i}}\right|^2\\&=\frac{1}{A_0\bar{d}_0^{k}a(f)^{\bar{d}_0}}\left|\sum_p\sum_i h_{p,i}e^{-j2\pi f\hat{\tau}_{p,i}}\right|^2\end{aligned} \tag{5-39}$$

其中，$\dot{\tau}_{p,i}=\tau_{p,i}-\vartheta_p\tilde{t}$，且第二个等式成立利用了公式（5-4）。考虑到不同传播路径长度、小尺度的散射以及多普勒频移相对较小，可以得到$0<f\dot{\tau}_{p,i}\le 1$。因此，式（5-39）可以近似为

$$\begin{aligned}|H(f)|^2 &\approx \frac{1}{A_0\bar{d}_0^k a(f)^{\bar{d}_0}}\left|\sum_p\sum_i h_{p,i}\right|^2\cos^2(2\pi f\bar{\tau})\\ &\approx \frac{1}{A_0\bar{d}_0^k a(f)^{\bar{d}_0}}\left|\sum_p\sum_i h_{p,i}\right|^2\left(1-4\pi^2 f^2\bar{\tau}^2\right)^2\end{aligned} \tag{5-40}$$

其中，第二个近似符号成立利用了余弦函数的泰勒级数展开。接着，对两边取对数和关于$\ln(1-x)$的泰勒级数展开，可得

$$\lg|H(f)|^2\approx\lg\frac{1}{A_0\bar{d}_0^k}+\lg a(f)^{-\bar{d}_0}+\lg\left|\sum_p\sum_i h_{p,i}\right|^2-\frac{8}{\ln 10}\pi^2\bar{\tau}^2 f^2 \tag{5-41}$$

对于式（5-38）的第二项，根据文献[141]，环境的噪声功率谱密度可以近似为

$$\lg N(f)\approx N_1-\eta\lg f \tag{5-42}$$

其中，N_1和η是需要估计的参数。

最后，结合式（5-39）、（5-41）和（5-42），可以得到频率和速率的线性关系

$$\lg\left(2^{r/B}-1\right)+\frac{8}{\ln 10}\pi^2\bar{\tau}^2 f^2-\eta\lg f+\lg a(f)^{\bar{d}_0}-\lg\left(\frac{P\left|\sum_p\sum_i h_{p,i}\right|^2}{A_0\bar{d}_0^k B10^{N_1}}\right)=0 \tag{5-43}$$

令$c_5=-\lg\left(P\left|\sum_p\sum_i h_{p,i}\right|^2\Big/A_0\bar{d}_0^k B10^{N_1}\right)$，$c_2=8\pi^2\bar{\tau}^2/\ln 10, c_3=-\eta$和$c_1=c_4=1$，可得

$$c_1\varphi_1(r)+c_2\varphi_2(f)+c_3\varphi_3(f)+c_4\varphi_4(f)+c_5=0 \tag{5-44}$$

其中，$\varphi_1(r)=\lg\left(2^{r/B}-1\right),\varphi_2(f)=f^2,\varphi_3(f)=\lg f,\varphi_4(f)=\lg a(f)^{\bar{d}_0}$。 □

下面利用数值仿真来验证上述模型。在该线性模型中，其系数和变量分别$\mathbf{w}=[c_1,c_2,c_3,c_4,c_5]$和$\mathbf{x}=[\varphi_1(r),\varphi_2(f),\varphi_3(f),\varphi_4(f),1]^\top$，则各动作上的估计的传输成功概率$\hat{\theta}_a$可以通过 Sigmoid 函数获得，即

$$\hat{\theta}_a(\mathbf{x}) = \frac{1}{1+e^{-\hat{\mathbf{w}}\mathbf{x}}} \tag{5-45}$$

其中，$\hat{\mathbf{w}}$ 表示需要学习的模型参数。通过观测 Sigmoid 函数的输出（即当传输速率高于当前信道最大容量时输出近似为 0；否则，输出近似为 1），可以判断当前动作属于哪个集合，从而达到分类的目的。与此同时，信道的真实传输成功概率 θ_a 可以利用文献[139]中的方法获得，具体步骤如算法 5.3 所示。最后，该分类模型的目标是最小化如下互熵函数

$$\min_{\mathbf{w}} \ J(\mathbf{w}) : -\sum_{t=1}^{T}\left(\theta_{I_a^t}\ln\left(\hat{\theta}_{I_a^t}\right)+\left(1-\theta_{I_a^t}\right)\ln\left(1-\hat{\theta}_{I_a^t}\right)\right) \tag{5-46}$$

其中，I_a^t 表示时隙 t 被选择的动作。由于上式是关于参数 $\mathbf{w}$ 的凸函数，其解可以通过随机梯度下降方法（Stochastic Gradient Descent, SGD）得到,即 $c_i = c_i - \gamma\partial J(\mathbf{w})/\partial c_i$, $\forall c_i \in \mathbf{w}$，其中 γ 是迭代步长。

图 5.6 给出了利用该分类模型得到的动作空间分类结果，其中传输频率和速率的变化范围分别为 $[0,\ 2]$kbps 和 $[0,\ 20]$kHz。在图 5.6 中，黑色实线表示真实边界，虚线表示学习到边界，红色“o”表示可能存在最优动作的集合，蓝色“+”表示不可能存在最优动作的集合。模型参数值为 c_1=-73.4444， c_2=-1.7227 , c_3= 317.29 c_4=1.6806 , c_5=-29.2266。从图中可以看到，学习到的边界非常接近真实的边界，总的分类准确率达到了 98.46%。

算法 5.3 产生真实的传输成功概率 θ_a

步骤 1： 初始化参数： $B, P, f, \bar{d}_0, A_0, s, w, k, \xi, \Psi_r^{\text{ref}}$

步骤 2: 执行 $p = 1, 2, \ldots$ 次以下步骤（多径路径数目）

步骤 3: 执行 $p = 1, 2, \ldots$ 次以下步骤（散射路径数目）

步骤 4: 计算时延 $\tau_{p,i} = (d_{p,i} - \bar{d}_0)/c$ 和多普勒因子 $\vartheta_p = v_p / c$

步骤 5: 利用公式（5-3）计算信道转移函数 $H(f)$

步骤 6： 利用公式（5-7）计算接收信噪比 $\Psi(f)$

步骤 7： 利用公式（5-8）得到成功传输概率 θ

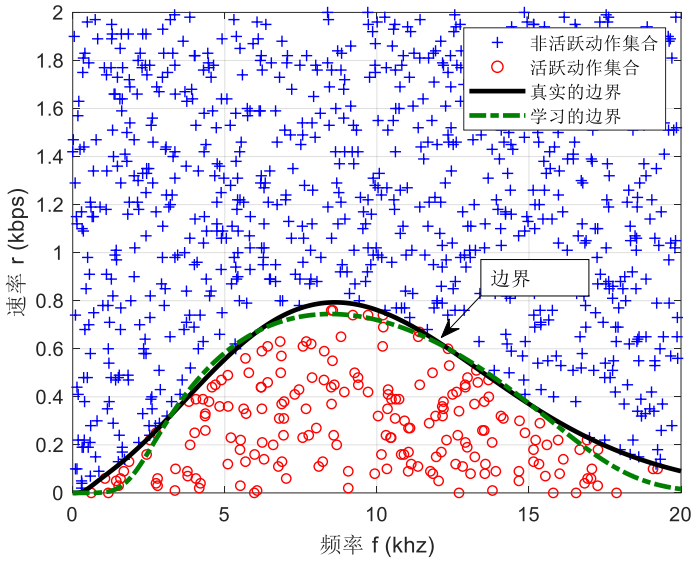


图 5.6 利用逻辑回归模型得到动作空间分类结果

算法 5.4 基于 IBS-TS 的选择策略

步骤 1： 初始化参数： $m,n,\sigma_f,\sigma_r,t_c,\mathcal{R},\mathcal{F},T$ ， 同时设定 $\hat{\mu}_a(0)=0,D_a^0=0,$ $\alpha_a=0,\beta_a=0,\hat{\theta}_a=1,\forall a\in\mathcal{A}$

步骤 2: 执行 $t=1,2,...,T$ 次以下步骤

TS 算法部分：

步骤 3: 利用 Beta 分布获取集合 $\mathcal{M}^+(L(t))$ 中所有动作的经验成功概率：

$$\tilde{\theta}_a \sim \text{Beta}(\alpha_a+1,\beta_a+1),\forall a\in\mathcal{M}^+(L(t))$$

步骤 5: 确定当前的动作： $I_a^t=\arg\max_{a\in\mathcal{A}}\left(r_a\tilde{\theta}_a\hat{\theta}_a\right)$

步骤 6： 利用动作 I_a^t 上的频率和速率进行传输，并观测传输反馈 $X_{I_a^t}(t)$

步骤 8： 更新动作 I_a^t 参数：当 $X_{I_a^t}(t)=1$ 时， $\alpha_{I_a^t}=\alpha_{I_a^t}+1$；否则， $\beta_{I_a^t}=\beta_{I_a^t}+1$

动作分类模型：

步骤 9： 当时隙数大于 t_c 时（即 $t\geq t_c$ ），执行以下步骤；否则，跳转至步骤 3

步骤 10： 利用 SGD 方法更新模型参数： $c_i=c_i-\gamma\partial J(\mathbf{w})/\partial c_i,\forall c_i\in\mathbf{w}$

步骤 11： 利用式（5-45）计算估计的成功传输概率 $\hat{\theta}_a,\forall a\in\mathcal{A}$

最后，基于上述逻辑分类模型，本节提出一种 IBS-TS 算法来求解问题（5-33），其伪代码由算法 5.4 给出。首先，算法初始化各动作上的估计的成功传输概率，即令 $\hat{\theta}_a = 1, \forall a \in \mathcal{A}$ 负责均匀探索。在每个时隙，根据平均吞吐量乘以估计的成功传输概率的值来选择当前时刻动作，即 $I_a^t = \arg\max_{a\in\mathcal{A}} \left(r_a \tilde{\theta}_a \hat{\theta}_a \right)$。其中，当时间足够长时，利用 Sigmoid 函数得到的估计的传输成功概率非 0 即 1。在这种情况下，算法只需探索部分的动作，而不是整个动作空间。在动作分类模型部分，为了降低计算复杂度，这里提前设定好一个时间门限 t_c，当时隙 t 超过 t_c 时才进行分类操作。具体地，首先利用 SGD 方法更新模型参数 $\mathbf{w}$；其次，利用 Sigmoid 函数产生估计的成功传输概率 $\hat{\theta}_a$，达到分类的目的。

图 5.7 给出了 IBS-TS 算法在不同时间门限 t_c 设置下，累积遗憾随时隙变化的曲线。其中，水声信道参数和分类模型参数由表 5.1 给出。从图中可以看到，所提算法在不同的门限下的累积遗憾都呈对数增长，这表明算法可以收敛到最佳动作；另外，门限 t_c 越小，算法的性能越好；但当 $t_c \leq 100$ 时，算法的性能增益不明显。因此，可以通过调整门限 t_c 来对算法的性能增益和计算复杂度进行权衡。

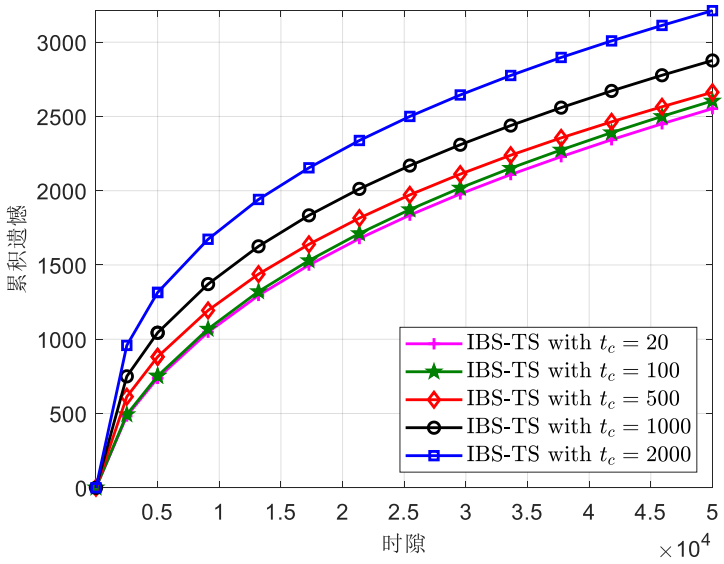


图 5.7 IBS-TS 算法在不同 t_c 下累积遗憾随时隙变化的曲线

5.4 理论分析

本节利用数理统计知识与有限时间分析技术推导所提算法的遗憾上界。考虑到 HCD-UO-TS 算法与 IBS-TS 算法都基于 TS 算法。因此，本章只推导 UO-TS 算法的遗憾上界。

定理 5.1. 令 μ^* 表示最佳动作 a^* 的平均奖励。对于任意的 $\epsilon>0$ 和给定的单峰图 $\mathcal{G}=(\mathcal{V},\mathcal{E})$，UO-TS 算法的伪遗憾的期望上界为

$$\mathcal{R}eg(T)\le(1+\epsilon)\sum_{a\in\mathcal{M}^+(a^*)}\frac{\Delta_a\ln T}{\mathrm{KL}(\mu_a,\mu^*)}+\mathcal{O}\left(\frac{|\mathcal{M}^+(a^*)|}{\epsilon^2}\right)+(mn-1)\left(\frac{2}{\Delta_{a'}^2}+\Delta_{a'}^2\right) \tag{5-47}$$

其中，$\mathrm{KL}\left(\mu_a,\mu^*\right)=\mu_a\ln\left(\mu_a/\mu^*\right)+(1-\mu_a)\ln\left((1-\mu_a)/(1-\mu^*)\right)$；$\ln(\cdot)$ 表示以自然数为底的对数；此外，$\Delta_a=\mu^*-\mu_a$ 和 $\Delta_{a'}=\max_{a\in\mathcal{M}^+(a)}\mu_a-\mu_{a'}$。

证明： 与传统的 TS 算法相比，UO-TS 算法主要有两个区别：第一，UO-TS 算法感兴趣的是最大化平均吞吐量 $r_a\theta_a$，而不只是成功传输概率 θ_a；第二，UO-TS 算法考虑了目标函数中的二维单峰特征，因此，其只需要在其邻居集合 $\mathcal{M}^+$ 上搜索最佳动作，而不是在整个动作空间 $\mathcal{A}$ 上进行。基于这两点，下面给出定理 5.1 的证明。

首先，根据式（5-12），UO-TS 算法的伪遗憾的期望可以写成

$$\begin{aligned}\mathcal{R}eg(T)&=\sum_{a_{i,j}\in\mathcal{A}}\Delta_{a_{i,j}}\mathbb{E}\left[D_{a_{i,j}}(T)\right]\\&=\sum_{a_{i,j}\in\mathcal{A}}\Delta_{a_{i,j}}\mathbb{E}\left[1\{I_t=a_{i,j}\}\right]\end{aligned} \tag{5-48}$$

其中，$1\{\cdot\}$代表指示函数。由于目标函数的单峰特性，动作选择存在两个情况：第一，leader 为最优动作的情况；第二，leader 为非最优动作的情况，即 $1\{I_t=a_{i,j}\}=1\{L(t)=a^*,I_t=a_{i,j}\}+1\{L(t)\neq a^*,I_t=a_{i,j}\}$。因此，上式可以重写为

$$\begin{aligned}\mathcal{R}eg(T) &= \sum_{a_{i,j}\in\mathcal{A}} \Delta_{a_{i,j}} \mathbb{E}\left[\sum_{t=1}^{T} 1\{I_t = a_{i,j}\}\right] \\ &= \underbrace{\sum_{a_{i,j}\in\mathcal{M}^+(a^*)} \Delta_{a_{i,j}} \mathbb{E}\left[\sum_{t=1}^{T} 1\{L(t)=a^*, I_t = a_{i,j}\}\right]}_{\mathcal{R}eg_1} \\ &\quad + \underbrace{\sum_{a_{i,j}\neq a^*} \Delta_{a_{i,j}} \mathbb{E}\left[\sum_{t=1}^{T} 1\{L(t)\neq a^*, I_t = a_{i,j}\}\right]}_{\mathcal{R}eg_2}\end{aligned} \tag{5-49}$$

对于第一项，其相当于在集合内 $\mathcal{M}^+(a^*)$ 运行 TS 算法，因此 $\mathcal{R}eg_1$ 可以利用文献[84]中的结果得到，即

$$\mathcal{R}eg_1 \le (1+\epsilon) \sum_{a\in\mathcal{M}^+(a^*)} \frac{\ln T}{\mathrm{KL}(\mu_a, \mu^*)} \Delta_a + C_1 \tag{5-50}$$

其中，$C_1 = \mathcal{O}(|\mathcal{M}^+(a^*)|/\epsilon^2)$。

对于第二项 $\mathcal{R}eg_2$，首先需要量化所有次优动作在时间 T 内被选择的次数，即等价于量化次优动作成为 leader 的概率。因此，第二项可以展开成

$$\begin{aligned}\mathcal{R}eg_2 &= \sum_{a_{i,j}\neq a^*} \Delta_{a_{i,j}} \mathbb{E}\left[\sum_{t=1}^{T} 1\{L(t)\neq a^*, I_t = a_{i,j}\}\right] \\ &\le \sum_{a_{i,j}\neq a^*} \mathbb{E}\left[\sum_{t=1}^{T} 1\{L(t)\neq a^*, I_t = a_{i,j}\}\right] \\ &\le \sum_{a_{i,j}\neq a^*} \mathbb{E}\left[l_{a_{i,j}}(T)\right]\end{aligned} \tag{5-51}$$

其中，$l_{a_{i,j}}(T)$ 表示动作 $a_{i,j}$ 在时间 T 内成为 leader 的次数，且第一个不等式成立是因为 $\Delta_{a_{i,j}} \le 1$。根据传统的随机 MAB 分析方法，上式可进一步转换成

$$\begin{aligned}\mathcal{R}eg_2 &\le \sum_{a_{i,j}\neq a^*} \sum_{t=1}^{T} \mathbb{E}\left[1\{L(t)=a_{i,j}\}\right] \\ &\le \sum_{a_{i,j}\neq a^*} \sum_{t=1}^{T} \mathbb{E}\left[1\left\{\hat{\mu}_{a_{i,j}}(t) = \max_{a'_{i,j}\in\mathcal{M}^+(a_{i,j})} \hat{\mu}_{a'_{i,j}}(t)\right\}\right] \\ &= \sum_{a_{i,j}\neq a^*} \sum_{t=1}^{T} \mathbb{E}\left[1\left\{\hat{\mu}_{a_{i,j}}(t) \ge \hat{\mu}_{a'_{i,j}}(t)\right\}\right] \\ &= \sum_{a_{i,j}\neq a^*} \sum_{t=1}^{T} \Pr\left(\hat{\mu}_{a_{i,j}}(t) \ge \hat{\mu}_{a'_{i,j}}(t)\right)\end{aligned} \tag{5-52}$$

其中，第一个不等式成立是因为考虑了估计的 leader，而不是真实的 leader；令 $\Delta_{a'_{i,j}}$ 表示选择 $a_{i,j}$ 而不是其邻居 $a'_{i,j}$ 的期望遗憾，则上式变成，

$$\begin{aligned}\mathcal{R}eg_2 &\le \sum_{a_{i,j}\neq a^*}\sum_{t=1}^{T}\Pr\left(\hat{\mu}_{a_{i,j}}(t)-\mu_{a_{i,j}}-\frac{\Delta_{a'_{i,j}}}{2}-\hat{\mu}_{a'_{i,j}}(t)+\mu_{a'_{i,j}}-\frac{\Delta_{a'_{i,j}}}{2}\ge 0\right)\\ &\le \sum_{a_{i,j}\neq a^*}\underbrace{\sum_{t=1}^{T}\Pr\left(\hat{\mu}_{a_{i,j}}(t)-\mu_{a_{i,j}}-\frac{\Delta_{a'_{i,j}}}{2}\ge 0\right)}_{\mathcal{R}eg_{21}}\\ &\quad+\sum_{a_{i,j}\neq a^*}\underbrace{\sum_{t=1}^{T}\Pr\left(\hat{\mu}_{a'_{i,j}}(t)-\mu_{a'_{i,j}}+\frac{\Delta_{a'_{i,j}}}{2}\le 0\right)}_{\mathcal{R}eg_{22}}\end{aligned} \tag{5-53}$$

对于上式第一项 $\mathcal{R}eg_{21}$，可以得到

$$\mathcal{R}eg_{21}=\sum_{t=1}^{T}\Pr\left(\hat{\mu}_{a_{i,j}}(t)-\mu_{a_{i,j}}\ge\frac{\Delta_{a'_{i,j}}}{2}\right)\le\sum_{t=1}^{T}\exp\left(\frac{-t\Delta^2_{a'_{i,j}}}{2}\right)\le\frac{2}{\Delta^2_{a'_{i,j}}}=C_2 \tag{5-54}$$

其中，第一个不等式成立是利用了 Hoeffding 不等式；第二个不等式成立是利用了积分 $\sum_{t=1}^{T}e^{-tx}\le\int_0^{+\infty}e^{-ux}du=1/x$ 的结果。同样，对于第二项有

$$\mathcal{R}eg_{22}=\sum_{t=1}^{T}\Pr\left(\left|\hat{\mu}_{a'_{i,j}}(t)-\mu_{a'_{i,j}}\right|\ge\frac{\Delta_{a'_{i,j}}}{2}\right)\le 2\sum_{t=1}^{T}\exp\left(\frac{-t\Delta^2_{a'_i,j}}{2}\right)\le\Delta^2_{a'_i,j}=C_3 \tag{5-55}$$

综上，UO-TS 算法的伪遗憾的期望上界为

$$\begin{aligned}\mathcal{R}eg(T)&\le\mathcal{R}eg_1+\sum_{a_i\neq a^*}\left(\mathcal{R}eg_{21}+\mathcal{R}eg_{22}\right)\\ &\le(1+\epsilon)\sum_{a\in\mathcal{M}^+(a^*)}\frac{\ln T}{\mathrm{KL}(\mu_a,\mu^*)}\Delta_a+C_1+(mn-1)(C_2+C_3)\end{aligned} \tag{5-56}$$

因此，定理 5.1 证毕。 □

注 1. 定理 5.1 表明 UO-TS 算法的遗憾上界随时间 T 呈对数增长。因此，当时间 T 足够大时，UO-TS 算法的每次迭代的遗憾将趋于 0。也就是说，此时链路传输在最优的频率和速率上。

注 2. 从定理 5.1 还可以看到，UO-TS 算法的遗憾上界主要依赖于最佳动作 a^* 的邻居动作集合 $\mathcal{M}\left(a^*\right)$。这表明，UO-TS 算法对动作空间的大小不敏感，而且只需

要在很小一部分的动作空间中探索，从而表现出比传统的随机 MAB 算法更好的性能，比如 UCB、KL-UCB 和 TS 等算法。此外，定理 5.1 还表明 UO-TS 算法比传统的 MAB 算法提高了约 $\log_2(MN/5)$ 倍，其中，5 和 MN 分别表示 UO-TS 算法和传统 MAB 算法需要探索的动作数目。

5.5 仿真结果

在仿真中，水声信道由算法 5.3 产生，其参数由表 5.1 给出。所有的结果都由 10^3 次蒙特卡洛仿真得到，算法的总迭代次数或时隙数为 $T=5\times10^4$。

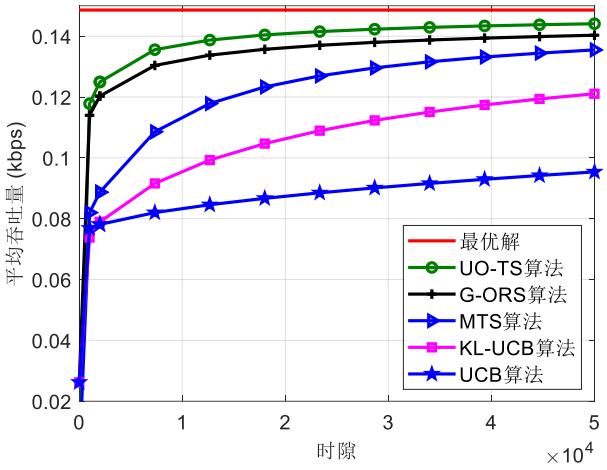
首先，考虑平稳信道的情形。假设水声链路可选择的频率和速率为 $\{4,6,8,10,14\}$ kHz 和 $\{0.2,0.4,0.6,0.9,1.2,1.4,1.6,2\}$ kbps；同时，考虑三种不同的成功传输概率模型，其对应的接收信噪比门限为

$$\begin{aligned}&(1.0,2.0,3.0,4.0,5.0,6.0,7.0,8.0)\times10^{-3} \quad \text{Case I}\\&(1.0,1.5,2.0,2.5,3.0,7.0,7.5,8.0)\times10^{-3} \quad \text{Case II}\\&(1.0,4.0,4.5,5.0,5.5,6.0,7.0,8.0)\times10^{-3} \quad \text{Case III}\end{aligned} \tag{5-57}$$

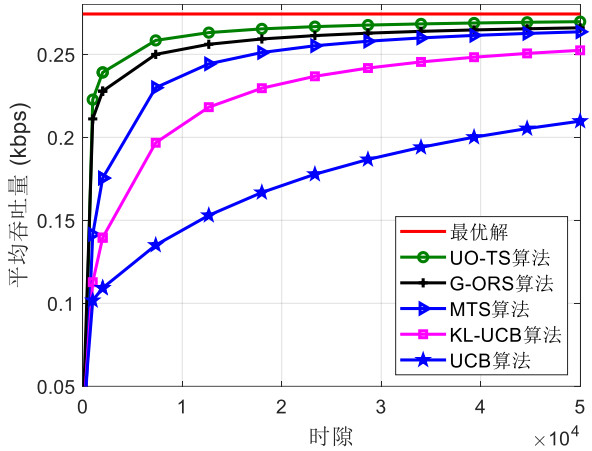
接着，利用算法 5.3 可以得到频率和速率的联合成功传输概率。

表 5.1 仿真参数设置

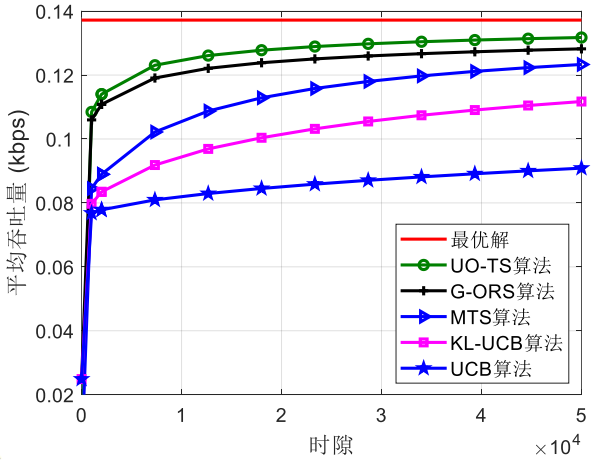
符号	物理意义	数值
$\bar{d}_0$	链路传输距离	5Km
P	发送功率	60 dBm
k	信号传播的几何形状	1.5
B	信号带宽	0.1KHz
w	水面风速	1
s	船舶活动因子	0
γ	ERWA 方法的权重因子	0.1
τ_p	路径 p 的相对时延	$[0,15,30,45,60,75]/1500$
ξ_p	路径 p 的累积反射系数	$[0.2,0.31,0.28,0.26,0.13,0.11]$



（a）Case I



（b）Case II



（c）Case III

图 5.8 UCB、KL-UCB、MTS、G-ORS 和 UO-TS 算法在 Case I、Case II 和 Case III 情况下的性能曲线

此外几种对比算法的设置具体如下：

- 最优解：当链路传输在最优频率和速率时，得到的平均吞吐量，其中，最优频率和速率通过上帝视角获得；
- MTS 算法：改进的汤普森采样算法。与本文所提 UO-TS 算法相比，MTS 算法没有利用目标函数的单峰特性，因此需要探索整个动作空间；其参数设置与 UO-TS 算法一致，但不需要确定 leader 这一步骤；
- UCB 算法：该算法由表 1.1 给出，即公式（1-7），其中参数$c=0.01$；
- KL-UCB 算法：同样该算法由表 1.1 给出，即公式（1-8），其中参数$c=0.01$；
- G-ORS 算法：该算法由文献[123]给出，可以看成是基于单峰特性的 KL-UCB 算法。即在 KL-UCB 算法的基础上，每回合确定 leader，然后在 leader 及其邻居动作上搜素最佳的动作。值得注意的是，KL-UCB 算法每回合需要调用牛顿法，因此具有较高的计算复杂度。

图 5.8 给出了 UCB、KL-UCB、MTS、G-ORS 和 UO-TS 算法在 Case I、Case II 和 Case III 情况下的性能。从图中可以看出，UT-OS 算法在三种情况下都拥有最好的性能，且接近最优解。然而，UCB、KL-UCB 和 MTS 算法的性能较差。其原因是它们需要探索整个动作空间，而 UT-OS 算法只需要探索最优动作的邻居动作集合。此外，G-ORS 算法虽然考虑了单峰特征，但因是 G-ORS 算法基于 KL-UCB 算法；在奖励分布先验已知的情况，TS 算法通常优于 KL-UCB 算法[55]。

其次，考虑非平稳信道的情况。对于突变信道，这里考虑两次突变情况，即$t=1.5\times10^4$和$t=3\times10^4$。在第一次突变点发生时，动作上的成功传输概率由 Case I 转移到 Case II;当第二次变点发生时，成功传输概率由 Case II 转移到 Case III。此外，对于缓变信道，其模型和参数设置与图 5.5 中相同，如动作的平均吞吐量每隔100个时隙在区间$[\mu_a-0.1,\mu_a+0.1]$内发生随机变化。为了方便比较，假设对比算法同样具有所提算法相同的联合检测非平稳信道的能力，即通过仿真来对比 HCD-UCB、HCD-KL-UCB、HCD-MTS 和 HCD-UO-TS 算法的性能。

图 5.9 给出了 HCD-UCB、HCD-KL-UCB、HCD-MTS 和 HCD-UO-TS 算法的累积遗憾随着时隙变化的曲线。从图中可以看到，GLR 检验方法可以有效地检测出两个突变点，即$t=1.5\times10^4$和3×10^4。当检测到突变点后，所有算法将重新初始

化其参数。然而，相同条件下，HCD-UO-TS 算法拥有最佳的性能，且收敛速度最快。图 5.9 表明 HCD-UO-TS 算法不仅可以有效处理非平稳通道（即突变和缓变信道），而且还有较好的性能。

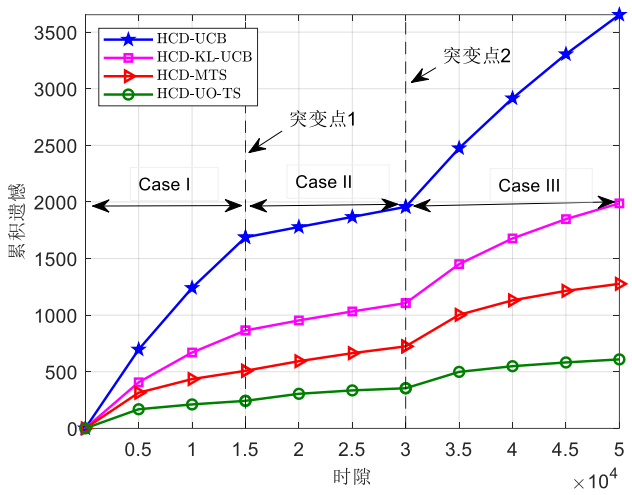


图 5.9 在非平稳信道下 HCD-UCB、HCD-KL-UCB、HCD-MTS 和 HCD-UO-TS 算法的累积遗憾随着时隙变化的曲线

最后，考虑动作空间较大、目标函不数存在单峰特性的情况。其中，传输速率和频率分别由 0 到 2 kbps 和由 0 到 20 kHz 变化。信噪比门限由公式 $\Psi_r^{\mathrm{ref}} = 10^{-3} + 10^{-2} / \left(1 + e^{-\varrho(r-1)}\right)$ 给定，其中 ϱ 是值在 $(0,1)$ 之间的一个尺度变量。因此，通过调整 ϱ 的大小可以改变传输成功概率的分布。当 ϱ 较小时，动作空间的传输成功概率均匀分布在 $(0,1)$；当 ϱ 较大时，较高速率的动作对应的传输成功概率都为 0，较低的都为 1，使得目标函数不具有单峰特性。此外，动作的真实传输成功概率由算法 5-3 得到。

当尺度参数为 $\varrho = 0.1$ 时，图 5.10 给出了 UCB、KL-UCB、MTS、G-ORS 和 UO-TS 算法的平均吞吐量随时隙变化的曲线。从图中可以看到，IBS-TS 算法和 UO-TS 算法的性能优于 UCB、KL-UCB 和 MTS 算法，并且接近最优解。此外，IBS-TS 算法的性能略低于 UO-TS 算法，这是因为 $\varrho = 0.1$ 时目标函数仍具有单峰特征。图 5.10 表明即使目标函数存在单峰特征，IBS-TS 算法也具有与 UO-TS 算法相当的性能。

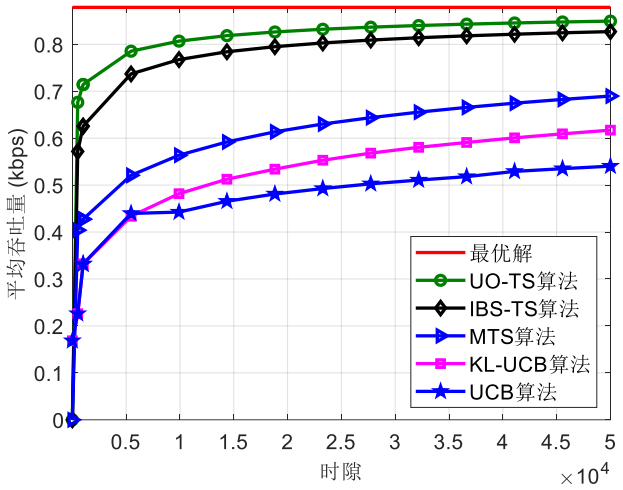


图 5.10 在大动作空间下 UCB、KL-UCB、MTS、G-ORS 和 UO-TS 算法的平均吞吐量随时隙变化的曲线，其中 $\varrho=0.1$

与之对比，图 5.11 给出的是 UCB、KL-UCB、MTS、G-ORS 和 UO-TS 算法的平均吞吐量随时隙变化的曲线，其中，尺度参数值设定为 $\varrho=1$。这种情况下，较高速率的动作对应的传输成功概率都为 0，较低的都为 1。从图中可以看到，IBS-TS 算法在对比算法中具有最好的性能，且收敛速度最快。然而，由于目标函数的单峰特征不成立，UO-TS 算法的性能最差。综上所述，图 5.10 和 5.11 表明无论目标函数中是否存在单峰特征 IBS-TS 算法都可以快速收敛到最佳动作。

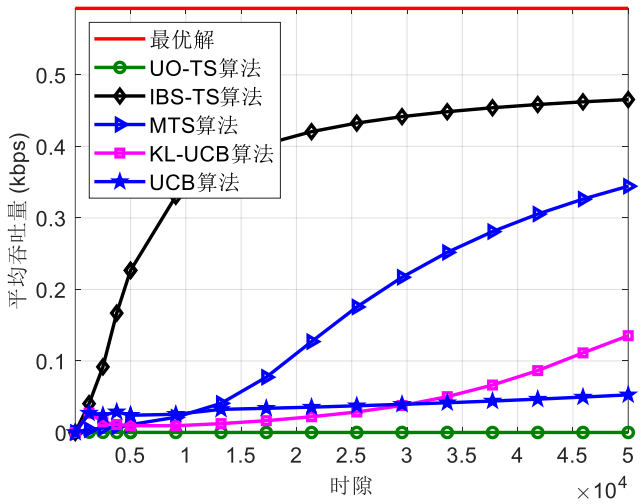


图 5.11 在大动作空间下 UCB、KL-UCB、MTS、G-ORS 和 UO-TS 算法的平均吞吐量随时隙变化的曲线，其中 $\varrho=1$

5.6 本章小结

本章研究了水声通信链路的传输频率和速率的联合选择问题。首先，基于在线学习理论将该问题建模成一个随机 MAB 框架，序贯地学习出最优的传输频率和速率。考虑到传统的随机 MAB 框架收敛速度较慢，进一步提出了一种基于模型的随机 MAB 框架。然后，分别考虑目标函数的二维单峰结构、水声信道的非平稳特点以及传输速率与频率之间关系，分别提出了 UO-TS、HCD-UO-TS 和 IBS-TS 算法来解决该联合选择问题。此外，针对 UO-TS 算法，推导了其累积遗憾的理论上界，并表明其收敛速度比传统的 MAB 算法提高了约 $\log_2(MN/5)$ 倍，其中 M 和 N 分别是链路上传输频率和速率的数目。最后，仿真结果验证了 UO-TS 算法理论结果，且表明所提算法优于当前最先进的随机 MAB 算法。而且，在动作空间的较大的场景下，所提算法的收敛速度不受影响，这将有利于算法的实际部署。

第 6 章 总结与展望

本章主要对前五章的研究内容进行总结，同时针对下一代无线通信网络中的资源分配问题与 MAB 技术的结合做进一步分析，并探索未来的可能研究方向。

6.1 全文总结

随着无线通信技术的快速发展，传统的资源分配与优化方法和技术已经无法满足下一代无线通信网络的需求和指标。本论文围绕下一代无线通信网络中的典型应用场景，针对 FD-CSMA 网络空间复用、分布式异构蜂窝网络资源分配、高密度物联网设备调度、水声通信网络链路自适应四个问题，结合优化理论、博弈理论、马尔科夫决策过程和数据驱动技术，提出了一系列基于 MAB 技术的资源分配与优化策略，为实现下一代无线通信网络的愿景与指标提供理论依据和技术支撑。主要内容和研究结果归纳如下：

（1）研究了 FD-CSMA 网络的空间复用问题，通过考虑网络中各条 FD 链路的 TP 控制、CST 调整和 LAI 自适应，提出了一种在随机和对抗环境中都最优的 FD-CSMA 算法。首先，基于 CSMA 的载波侦听模型和连续马尔科夫模型将该问题建模成一个传统的网络优化问题。然后，利用分解理论将其分解为 MAC 层的调度问题和物理层的参数选择问题。针对 MAC 层的调度问题，提出了一种最优的 FD-CSMA 算法得到最优 LAI；针对物理层的参数选择问题，提出了一种在随机和对抗环境中均最优的 MAB 算法得到最佳的 TP 和 CST 组合。最后，通过交替求解这两个子问题来近似求解该网络吞吐量最大化问题。此外，理论分析了所提算法的伪遗憾上界，且数值结果表明，与随机选择方法相比，所提算法的性能在静态和动态网络场景中分别提高了 48% 和 43% 。

（2）研究了分布式异构蜂窝网络中关于 IoT 设备的 RIS 和 SF 联合分配问题，提出了一种完全分布式的资源分配策略。首先，利用不同 SF 之间的单调特性将该问题建模成一个两阶段的 MPMAB 问题。然后，针对第一阶段的 MPMAB 问题，结合 ϵ-贪婪算法和非合作博弈方法，提出了一种最优 RIS 分配策略；针对第二阶段的 MPMAB 问题，结合 TS 算法提出了一种最优 SF 分配策略。最后，通过交替

迭代求解第一和第二阶段 MPMAB 问题，提出了一种完全分布式的 E2Boost 算法来求解该联合分配问题。此外，理论分析了所提算法的伪遗憾上界，即 $\mathcal{O}(\log_2^{1+\delta} T)$，表明当时间 T 足够大时，算法每回合的遗憾将趋向于 0。最后，仿真结果验证所提算法的遗憾上界，且其性能优于已有的分布式分配算法；而且，由于 E2Boost 算法采用了两阶段分配的机制，其收敛速度对动作空间的大小不敏感。

（3）研究了高密度物联网中基于信息新鲜度的 IoT 设备调度问题，提出了一种基于 WI 的调度策略，最小化网络的平均信息新鲜度。首先，考虑 IoT 设备之间的相关性将该设备调度问题建模成一个 MDP 问题。其次，为了降低计算复杂度，进一步将其建模为一个 CRMAB 问题。然后，考虑理想和非理想的信道模型，分别推导了 GWI 和 GPWI 的闭式表达式，提出了基于 GWI 和 GPWI 的调度策略来最小化网络的平均信息新鲜度。此外，通过求解松弛的拉格朗日问题，理论分析了所提策略的性能下界。最后，利用数值结果验证了该理论分析结果，并表明所提策略的性能优于已有的调度策略。而且，在高密度网络场景中所提策略可以显著降低网络的平均信息新鲜度。

（4）研究了单条水声通信链路的传输频率和速率的联合选择问题，提出了一类收敛速度快、易于实现和严格理论性能保证的算法来提高链路的传输效率。首先，考虑到水声信道的状态信息无法准确获取，将其建模成一个随机 MAB 框架来依次学习出最优的频率和速率。为了提高学习速率，利用问题模型的特性进一步提出了一种基于模型的随机 MAB 框架。接着，考虑目标函数的二维单峰结构、水声信道的非平稳特征以及传输频率与速率之间的关系，分别提出了 UO-TS 算法、HCD-UO-TS 算法和 IBS-TS 算法来解决该随机 MAB 问题。此外，理论分析了 UO-TS 算法的伪遗憾上界，并表明其收敛速度比传统的 MAB 算法提高了约 $\log_2(MN/5)$ 倍，其中 M 和 N 分别是传输频率和速率的数目。最后，仿真结果表明所提算法优于当前最先进的 MAB 算法，且其收敛速率对动作空间的大小不敏感。

6.2 后续工作展望

本论文研究了基于 MAB 技术的无线网络资源分配与优化策略，考虑了 FD-CSMA 网络空间复用、分布式异构蜂窝网络资源分配、高密度物联网设备调度、水

声通信网络链路自适应四个问题，下面将从如下几个方面开展未来的研究工作：

（1）基于对抗 MAB 的全双工 CSMA 网络空间复用机制

在前期工作中，本文假设 CSMA 的载波侦听机制是理想的，即全双工链路之间的传输不会发生碰撞。在这种情况下，网络的吞吐量可以利用连续时间马尔科夫模型求解得到，进而采用传统的优化理论中的次梯度下降方法获得最佳的 LAI 参数。但是，当考虑非理想的 CSMA 侦听机制时，网络中将存在隐藏终端和暴露终端问题，从而导致复杂的网络吞吐量计算模型（或连续时间马尔科夫模型失效）。针对这个问题，本文后续工作将深入分析 CSMA 协议中的侦听和退避机制，结合 MAB 技术联合考虑 MAC 层和物理层的参数选择问题，从全网的角度出发提出一种可靠、最优的全双工 CSMA 网络空间复用机制。

（2）基于 MPMAB 的分布式异构网络资源分配策略

在系统模型中，本文假设不同的 RIS 使用不同的频率，且一个 RIS 只能服务一个物联网设备。然而，在实际系统中，RIS 通常可以复用这些频率并服务于多个物联网设备。因此，本文的后续工作将建立在前期算法基础上对现有系统模型进行改进。一方面，设计一种可以使得蜂窝网用户和物联网设备的信号在同一个 RIS 中共存的机制；另一方面，考虑如何通过估计 RIS 和信道的准确状态信息来设计一种 RIS 辅助的多个物联网设备的机制，即 RIS 可以同时服务多个物联网设备，从而有效提高 RIS 的利用率。

（3）基于马尔科夫 MAB 和信息新鲜度的物联网设备调度策略

在前期工作中，本文通过为每一台 IoT 设备引入一个中间状态变量，使得 N 维的 MDP 问题可以解耦成 N 个一维的 MAB 子问题，从而有效降低问题的计算复杂度。但是，由于该中间状态变量无法完全刻画出 IoT 设备之间的相关性，解耦后的 MAB 问题的解和原问题的解之间存在一定的性能差距。针对这个问题，后续工作有两个思路。第一，利用最新的机器学习方法直接求解该 N 维的 MDP 问题。其难点在于如何求解一个状态空间连续的 MDP 问题。现有的方法通常对状态空间进行离散化或值函数近似，但这样只能得到的是原问题的近似解。因此，有必要对状态空间连续的 MDP 问题进行研究，进而将该结果作为本文所提算法的性能基准（即得到所提算法的数值遗憾上界，来验证理论分析结果）。第二，考虑在不引入

中间状态的情况下，如何将该 N 维的 MDP 问题可以解耦成 N 个一维的 MAB 子问题，然后采用基于 Whittle 索引的调度策略求解。其难点在于如何确定每台 IoT 设备上的信息新鲜度变化过程，因为一台设备的信息新鲜度更新不仅与其自身相关，还和它相关的设备的更新过程有关。因此，需要寻找一种合理的解耦方法在保证网络性能的前提下，同时降低算法的复杂度。这两种思路相辅相成，即思路一的求解可以为思路二的性能提供一个理论界上界；而思路二的求解又可以为思路一的求解过程的复杂度和理论分析提供参考。

（4）基于随机 MAB 的水声通信链路自适应机制

在系统模型中，本文只考虑单条水声通信链路上的自适应传输问题，忽略了链路之间干扰和调度问题。在后续工作中将考虑多条链路的水声通信网络，如何基于 MAB 技术通过传输频率和速率的选择来进一步提高该网络吞吐量？在这种情况下，需要仔细考虑水声的 MAC 协议，并通过联合优化 MAC 层参数和物理层参数来最大化网络吞吐量。由于水声传播速度低，传输时延对系统目标的影响需要仔细考虑。若将时延考虑在约束条件中，则可以利用该特性排除动作空间中明显次优的动作，而持续追踪最优的动作。若将时延考虑在目标函数中，则需要对奖励进行精心设计，使其能反映出时延对目标的具体影响，从而设计相应的 MAB 算法。

此外，如何将下一代无线通信网络中的资源分配与优化问题与 MAB 技术更好地结合在一起？进一步探索问题模型与算法设计的深入交互的可能性。如何从实际的资源分配与优化问题中提炼出新的 MAB 模型？进而为该模型设计相应的算法，并理论推导出所提算法的遗憾上界是本课题在未来的研究中重要且具有挑战性的一类问题。

参考文献

[1] Campolo C, Molinaro A, Iera A, et al. 5G network slicing for vehicle-to-everything services[J]. IEEE Wireless Communications, 2018, 24(6):38-45.

[2] Yang M, Li Y, Li B, et al. Service oriented 5G network architecture: an end-to-end software defining approach[J]. International Journal of Communication Systems, 2016, 29(10):1645-1657.

[3] Akyildiz I F, Kak A, Nie S. 6G and beyond: the future of wireless communications systems[J]. IEEE Access, 2020(99):133995-134030.

[4] 柴蓉, 邹飞, 刘莎,等. 6G 移动通信:愿景,关键技术和系统架构[J]. 重庆邮电大学学报：自然科学版, 2021, 33(3):1-11.

[5] 赵亚军, 郁光辉, 徐汉青. 6G 移动通信网络:愿景,挑战与关键技术[J]. 中国科学：信息科学, 2019, (8):1-25.

[6] Pogaku A C, Do D T, et al. UAV-assisted RIS for future wireless communications: a survey on optimization and performance analysis. IEEE Access, 2022, 10:16320-16336.

[7] You X, Wang X, Huang J, Gao X, et al. Towards 6G wireless communication networks: Vision, enabling technologies, and new paradigm shifts[J]. 中国科学：信息科学（英文版）, 2021, 64(1):74.

[8] Rajatheva N, Atzeni I, Bjornson E, et al. White paper on broadband connectivity in 6G[J]. arXiv:2004.14247, 2020.

[9] Tomkos I, Klonidis D, Pikasis E, et al. Toward the 6G network era: Opportunities and challenges[J]. IT Professional, 2020, 22(1):34-38.

[10] Liu G, Huang Y, Li N, Dong J, et al. Vision, requirements and network architecture of 6G mobile network beyond 2030[J]. China Communications, 2020, 17(9):92-104.

[11] Zhu G, Liu D, Du Y, et al. Toward an intelligent edge: Wireless communication meets machine learning[J]. IEEE Communications Magazine, 2020, 58(1): 19-25.

[12] Morocho-Cayamcela M E, Lee H, Lim W. Machine learning for 5G/B5G mobile and wireless communications: Potential, limitations, and future directions[J]. IEEE Access, 2019, 7:137184-137206.

[13] Jagannath J, Polosky N, Jagannath A, et al. Machine learning for wireless communications in the Internet of Things: A comprehensive survey[J]. Ad Hoc Networks, 2019, 93:1-97.

[14] Cunningham P, Cord M, Delany S J. Supervised learning in machine learning techniques for multimedia[M]. Springer, 2008.

[15] Barlow H B. Unsupervised learning[J]. Neural Computation, 1989, 1(3):295-311.

[16] Kaelbling L P, Littman M L, Moore A W. Reinforcement learning: A survey[J]. Journal of Artificial Intelligence Research, 1996, 4:237-285.

[17] Deng Y, Bao F, Kong Y, et al. Deep direct reinforcement learning for financial signal representation and trading[J]. IEEE Transactions on Neural Networks and Learning Systems, 2006, 28(3): 653-664.

[18] Carpi F, Häger C, Martalò M, et al. Reinforcement learning for channel coding: Learned bit-flipping decoding[C]. IEEE Annual Allerton Conference on Communication, Control, and Computing (Allerton), 2019: 922-929.

[19] Luong N C, Hoang D T, Gong S, et al. Applications of deep reinforcement learning in communications and networking: A survey[J]. IEEE Communications Surveys & Tutorials, 2019, 21(4):3133-3174.

[20] Tan J, Liang Y C, Zhang L, et al. Deep reinforcement learning for joint channel selection and power control in D2D networks. IEEE Transactions on Wireless Communications[J]. 2020, 20(2), 1363-1378.

[21] Ye H, Li G Y, Juang B H F. Deep reinforcement learning based resource allocation for V2V communications[]J. IEEE Transactions on Vehicular Technology, 2019, 68(4):3163-3173.

[22] Chu M, Li H, Liao X, et al. Reinforcement learning-based multiaccess control and battery prediction with energy harvesting in IoT systems[J]. IEEE Internet of Things Journal, 2018, 6(2):2009-2020.

[23] Sutton, Richard S, Andrew G B. Reinforcement learning: An introduction[M]. MIT Press, 2018.

[24] Lai T L, Robbins H. Asymptotically efficient adaptive allocation rules[J]. Advances in Applied Mathematics, 1985, 6(1): 4-22.

[25] Li F, Yu D, Yang H, et al. Multi-armed-bandit-based spectrum scheduling algorithms in wireless networks: A survey[J]. IEEE Wireless Communications, 2020, 27(1):24-30.

[26] Darak S J, Hanawal M K. Multi-player multi-armed bandits for stable allocation in heterogeneous ad-hoc networks[J]. IEEE Journal on Selected Areas in Communications, 2019 37(10):2350-2363.

[27] Setareh M, Hossain E. Multi-armed bandits with application to 5G small cells[J]. IEEE Wireless Communications, 2016, 23(3):64-73.

[28] Li X, Liu J, Yan L, Han S, et al. Relay selection for underwater acoustic sensor networks: A multi-user multi-armed bandit formulation[J]. IEEE Access, 2018, 6:7839-7853.

[29] Nikhil G, Dandekar K R. Learning state selection for reconfigurable antennas: A multi-armed bandit approach[J]. IEEE Transactions on Antennas and Propagation, 2013, 62(3):1027-1038.

[30] Xia W, Quek T Q, Guo K, et al. Multi-armed bandit-based client scheduling for federated learning[J]. IEEE Transactions on Wireless Communications, 2020, 19(11):7108-7123.

[31] 郭超平. 基于价格理论的无线网络资源分配算法研究[D]. 西安电子科技大学, 2014.

[32] 李超. 毫米波无线通信系统中的资源优化问题研究[D]. 浙江大学, 2019.

[33] 赵新胜, 尤肖虎. 未来移动通信系统中的无线资源管理[J]. 中兴通讯技术, 2002, 8(6): 7-10.

[34] Larsson E G, Edfors O, Tufvesson F, et al. Massive MIMO for next generation wireless systems[J]. IEEE Communications Magazine, 2014, 52(2): 186-195.

[35] Wang B, Liu K J R. Advances in cognitive radio networks: A survey[J]. IEEE Journal of Selected Topics in Signal Processing, 2010, 5(1): 5-23.

[36] Stuber G L, Barry J R, Mclaughlin S W, et al. Broadband MIMO-OFDM wireless communications[J]. Proceedings of the IEEE, 2004, 92(2): 271-294.

[37] Molisch A F, Ratnam V V, Han S, et al. Hybrid beamforming for massive MIMO: A survey[J]. IEEE Communications Magazine, 2017, 55(9): 134-141.

[38] Bharadia D, McMilin E, Katti S. Full duplex radios[C]. ACM SIGCOMM conference on SIGCOMM, 2013: 375-386.

[39] Lin S, Kernighan B W. An effective heuristic algorithm for the traveling-salesman problem[J]. Operations Research, 1973, 21(2): 498-516.

[40] Baccelli F, Klein M, Lebourges M, et al. Stochastic geometry and architecture of communication networks[J]. Telecommunication Systems, 1997, 7(1): 209-227.

[41] Han Z, Niyato D, Saad W, et al. Game theory in wireless and communication networks: theory, models, and applications[M]. Cambridge University Press, 2012.

[42] Jordan M I, Mitchell T M. Machine learning: Trends, perspectives, and prospects[J]. Science, 2015, 349(6245): 255-260.

[43] Dorigo M, Birattari M, Stutzle T. Ant colony optimization[J]. IEEE Computational Intelligence Magazine, 2006, 1(4): 28-39.

[44] Kennedy J, Eberhart R C. A discrete binary version of the particle swarm algorithm[C]. IEEE International Conference on Systems, man, and Cybernetics. Computational Cybernetics and Simulation, 1997, 5: 4104-4108.

[45] Whitley D. A genetic algorithm tutorial[J]. Statistics and Computing, 1994, 4(2): 65-85.

[46] Fang L. A generalized DEA model for centralized resource allocation[J]. European Journal of Operational Research, 2013, 228(2): 405-412.

[47] Lakshmanan H, De Farias D P. Decentralized resource allocation in dynamic networks of agents[J]. SIAM Journal on Optimization, 2008, 19(2): 911-940.

[48] Tong J, Fu L, Han Z. Throughput enhancement of full-duplex CSMA networks using multiplayer bandits[J]. IEEE Internet of Things Journal, 2021, 8(15): 11807-11821.

[49] Lin X, Shroff N B, Srikant R. A tutorial on cross-layer optimization in wireless networks[J]. IEEE Journal on Selected areas in Communications, 2006, 24(8): 1452-1463.

[50] Bubeck S, Stoltz G, Yu J Y. Lipschitz bandits without the lipschitz constant[C]. International Conference on Algorithmic Learning Theory. Springer, Berlin, Heidelberg, 2011: 144-158.

[51] Slivkins A. Multi-armed bandits on implicit metric spaces[J]. Advances in Neural Information Processing Systems, 2011, 24: 1-9.

[52] 徐勇军, 李国权, 徐鹏，等. 异构无线网络资源分配算法研究综述[J]. 重庆邮电大学学报: 自然科学版, 2018, 30(3):289-299.

[53] 谭伟, 张文新, 马雨出. LTE 的无线资源管理[J]. 邮电设计技术, 2007, (3): 61-64.

[54] Aleksandrs S. Introduction to multi-armed bandits[J]. arXiv preprint arXiv:1904.07272, 2019.

[55] Sébastien B, Nicolo C B. "Regret analysis of stochastic and nonstochastic multi-armed bandit problems[J]." Machine Learning, 2012, 5(1):1-122.

[56] Kuleshov, V, Doina P. Algorithms for multi-armed bandit problems[J]. arXiv preprint arXiv:1402.6028, 2014.

[57] Tran-Thanh L, Chapman A, De Cote E M, et al. Epsilon–first policies for budget–limited multi-armed bandits[C]. AAAI Conference on Artificial Intelligence. 2010.

[58] Auer P, Nicolo C B, Paul F. Finite-time analysis of the multiarmed bandit problem[J]. Machine Learning, 2020, 47(2):235-256.

[59] Garivier A, Cappé O. The KL-UCB algorithm for bounded stochastic bandits and beyond[C]. JMLR Workshop and Conference Proceedings, 2011: 359-376.

[60] Agrawal S, Goyal N. Further optimal regret bounds for thompson sampling[C]. Artificial Intelligence and Statistics. PMLR, 2013: 99-107.

[61] Auer P, Nicolo C B, Freund Y, et al. The non-stochastic multi-armed bandit problem[J]. SIAM Journal on Computing, 2002, 32(1):48-77.

[62] Gittins J C. Bandit processes and dynamic allocation indices[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1979, 41(2): 148-164.

[63] Whittle P. Restless bandits: Activity allocation in a changing world[J]. Journal of Applied Probability, 1988, 25(A): 287-298.

[64] Chu W, Li L, Reyzin L, et al. Contextual bandits with linear payoff functions[C]. JMLR Workshop and Conference Proceedings, 2011: 208-214.

[65] Zimmert J, Seldin Y. An optimal algorithm for stochastic and adversarial bandits[C]. International Conference on Artificial Intelligence and Statistics. PMLR, 2019: 467-475.

[66] Rosenski J, Shamir O, Szlak L. Multi-player bandits:A musical chairs approach[C]. International Conference on Machine Learning. PMLR, 2016: 155-163.

[67] Bistritz I, Leshem A. Distributed multi-player bandits: A game of thrones approach[J]. Advances in Neural Information Processing Systems, 2018, 31.

[68] Aurélien G, Lattimore T, Kaufmann E. On explore-then-commit strategies[J]. Advances in Neural Information Processing Systems, 2016, 29.

[69] Aurélien G, Moulines E. On upper-confidence bound policies for non-stationary bandit problems[J]. arXiv preprint arXiv:0805.3415, 2008.

[70] Trovo F, Paladino S, Restelli M, et al. Sliding-window thompson sampling for non-stationary settings[J]. Journal of Artificial Intelligence Research, 2020, 68:311-364.

[71] Cavenaghi E, Sottocornola G, Stella F, et al. Non stationary multi-armed bandit: Empirical evaluation of a new concept drift-aware algorithm[J]. Entropy, 2021, 23(3):380.

[72] Wei L, Srivastava V, Nonstationary stochastic multi-armed bandits: UCB policies and minimax regret[J], arXiv preprint arXiv:2101.08980, 2021.

[73] Liu F, Lee J, Shroff N. A change-detection based framework for piecewise-stationary multi-armed bandit problem[C]. AAAI Conference on Artificial Intelligence. 2018, 32(1).

[74] Ghatak G, Mohanty H, Rahman A U. Kolmogorov Smirnov test-based actively-adaptive thompson sampling for non-stationary bandits[J]. IEEE Transactions on Artificial Intelligence, 2021.

[75] Besson L, Kaufmann E. The generalized likelihood ratio test meets klucb: an improved algorithm for piece-wise non-stationary bandits[J]. Proceedings of Machine Learning Research, 2019, 1: 35.

[76] Chen W, Wang Y, Yuan Y. Combinatorial multi-armed bandit: General framework and applications[C]. PMLR, 2013, 151-159.

[77] Chen W, Hu W, Li F, et al. Combinatorial multi-armed bandit with general reward functions[J]. Advances in Neural Information Processing Systems, 2016, 29.

[78] Combes R, Proutiere A, Unimodal bandits: Regret lower bounds and optimal algorithms[J], arXiv preprint arXiv:1405.5096v1, 2014.

[79] Paladino S, Trovo F, Restelli M, et al. Unimodal thompson sampling for graph-structured arms[C]. AAAI Conference on Artificial Intelligence. 2017, 31(1).

[80] Combes R, Proutiere A, Yun D, et al. Optimal rate sampling in 802.11 systems[C]. IEEE INFOCOM, 2014: 2760-2767

[81] Buccapatnam S, Liu F, Eryilmaz A. Reward maximization under uncertainty: Leveraging side-observations on networks[J]. Journal of Machine Learning Research, 2018, 18:1–34.

[82] Combes R, Ok J, Proutiere A, et al. Optimal rate sampling in 802.11 systems: Theory, design, and implementation[J]. IEEE Transactions on Mobile Computing, 2018, 18(5): 1145-1158.

[83] Qi H, Hu Z, Wen X, et al. Rate adaptation with Thompson sampling in 802.11 ac WLAN[J]. IEEE Communications Letters, 2019, 23(10): 1888-1892.

[84] Gupta H, Eryilmaz A, Srikant R. Low-complexity, low-regret link rate selection in rapidly-varying wireless channels[C]. IEEE INFOCOM, 2018: 540-548.

[85] Gupta H, Eryilmaz A, Srikant R. Link rate selection using constrained thompson sampling[C]. IEEE INFOCOM, 2019: 739-747.

[86] Wilhelmi F, Barrachina-Munoz S, Bellalta B, et al. Potential and pitfalls of multi-armed bandits for decentralized spatial reuse in wlans[J]. Journal of Network and Computer Applications, 2019, 127: 26-42.

[87] Wilhelmi F, Cano C, Neu G, et al. Collaborative spatial reuse in wireless networks via selfish multi-armed bandits[J]. Ad Hoc Networks, 2019, 88: 129-141.

[88] Bubeck S, Slivkins A. The best of both worlds: Stochastic and adversarial bandits[C]. JMLR Workshop and Conference Proceedings, 2012: 42.1-42.23.

[89] Seldin Y, Slivkins A. One practical algorithm for both stochastic and adversarial bandits[C]. International Conference on Machine Learning. PMLR, 2014: 1287-1295.

[90] Audibert J Y, Bubeck S. Minimax policies for adversarial and stochastic bandits[C]. COLT. 2009, 7: 1-122.

[91] Zimmert J, Seldin Y. An optimal algorithm for stochastic and adversarial bandits[C]. International Conference on Artificial Intelligence and Statistics. PMLR, 2019: 467-475.

[92] Kadota I, Sinha A, Uysal-Biyikoglu E, et al. Scheduling policies for minimizing age of information in broadcast wireless networks[J]. IEEE/ACM Transactions on Networking, 2018, 26(6): 2637-2650.

[93] Jiang Z, Krishnamachari B, Zhou S, et al. Can decentralized status update achieve universally near-optimal age-of-information in wireless multiaccess channels[C]. IEEE International Teletraffic Congress (ITC 30), 2018, 1: 144-152.

[94] Maatouk A, Kriouile S, Assad M, et al. On the optimality of the Whittle's index policy for minimizing the age of information[J]. IEEE Transactions on Wireless Communications, 2020, 20(2): 1263-1277.

[95] Zou Y, Kim K T, Lin X, et al. Minimizing Age-of-Information in Heterogeneous Multi-Channel Systems: A New Partial-Index Approach[C]. ACM MOBIHOC, 2021: 11-20.

[96] Ta D T, Khawam K, Lahoud S, et al. Lora-mab: Toward an intelligent resource allocation approach for lorawan[C]. IEEE GLOBECOM, 2019: 1-6.

[97] Tibrewal H, Patchala S, Hanawal M K, et al. Distributed learning and optimal assignment in multiplayer heterogeneous networks[C]. IEEE INFOCOM, 2019: 1693-1701.

[98] Zafaruddin S M, Bistritz I, Leshem A, et al. Distributed learning for channel allocation over a shared spectrum[J]. IEEE Journal on Selected Areas in Communications, 2019, 37(10): 2337-2349.

[99] Bistritz I, Leshem A. Distributed multi-player bandits: A game of thrones approach[J]. Advances in Neural Information Processing Systems, 2018, 31:1-11.

[100] Liao Y, Wang T, Song L, et al. Listen-and-talk: Protocol design and analysis for full-duplex cognitive radio networks[J]. IEEE Transactions on Vehicular Technology, 2016, 66(1): 656-667.

[101] Everett E, Sahai A, Sabharwal A. Passive self-interference suppression for full-duplex infrastructure nodes[J]. IEEE Transactions on Wireless Communications, 2014, 13(2): 680-694.

[102] Boyd S, Boyd S P, Vandenberghe L. Convex optimization[M]. Cambridge University Press, 2004.

[103] Liu J, Yi Y, Proutiere A, et al. Towards utility‐optimal random access without message passing[J]. Wireless Communications and Mobile Computing, 2010, 10(1): 115-128.

[104] Liew S C, Kai C H, Leung H C, et al. Back-of-the-envelope computation of throughput distributions in CSMA wireless networks[J]. IEEE Transactions on Mobile Computing, 2010, 9(9): 1319-1331.

[105] Nevel'son M B, Has' minskii R Z. Stochastic approximation and recursive estimation[M]. American Mathematical Soc., 1976.

[106] Borkar V S. Stochastic approximation with ‘controlled Markov’ noise[J]. Systems & Control Letters, 2006, 55(2): 139-145.

[107] Wu Q, Zhang R. Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network[J]. IEEE Communications Magazine, 2019, 58(1): 106-112.

[108] ElMossallamy M A, Zhang H, Song L, et al. Reconfigurable intelligent surfaces for wireless communications: Principles, challenges, and opportunities[J]. IEEE Transactions on Cognitive Communications and Networking, 2020, 6(3): 990-1002.

[109] Abeywickrama S, Zhang R, Wu Q, et al. Intelligent reflecting surface: Practical phase shift model and beamforming optimization[J]. IEEE Transactions on Communications, 2020, 68(9): 5849-5863.

[110] Tong J, Jin M, Guo Q, et al. Cooperative spectrum sensing: A blind and soft fusion detector[J]. IEEE Transactions on Wireless Communications, 2018, 17(4): 2726-2737.

[111] Waret A, Kaneko M, Guitton A, et al. LoRa throughput analysis with imperfect spreading factor orthogonality[J]. IEEE Wireless Communications Letters, 2018, 8(2): 408-411.

[112] Menon A, Baras J S. Convergence guarantees for a decentralized algorithm achieving Pareto optimality[C]. IEEE American Control Conference, 2013: 1932-1937.

[113] Thomas M, Joy A T. Elements of information theory[M]. Wiley-Interscience, 2006.

[114] Zhang H, Di B, Song L, et al. Reconfigurable intelligent surfaces assisted communications with limited phase shifts: How many phase shifts are enough?[J]. IEEE Transactions on Vehicular Technology, 2020, 69(4): 4498-4502.

[115] Di B, Zhang H, Song L, et al. Hybrid beamforming for reconfigurable intelligent surface based multi-user communications: Achievable rates with limited discrete phase shifts[J]. IEEE Journal on Selected Areas in Communications, 2020, 38(8): 1809-1822.

[116] Lin J, Yu W, Zhang N, et al. A survey on Internet of Things: Architecture, enabling technologies, security and privacy, and applications[J]. IEEE Internet of Things Journal, 2017, 4(5): 1125-1142.

[117] Chen H, Gu Y, Liew S C. Age-of-information dependent random access for massive IoT networks[C]. IEEE INFOCOM WKSHPS, 2020: 930-935.

[118] Abd-Elmagid M A, Pappas N, Dhillon H S. On the role of age of information in the Internet of Things[J]. IEEE Communications Magazine, 2019, 57(12): 72-77.

[119] Kosta A, Pappas N, Angelakis V. Age of information: A new concept, metric, and tool[J]. Foundations and Trends in Networking, 2017, 12(3): 162-259.

[120] Qian Z, Wu F, Pan J, et al. Minimizing age of information in multi-channel time-sensitive information update systems[C]. IEEE INFOCOM, 2020: 446-455.

[121] Kadota I, Sinha A, Uysal-Biyikoglu E, et al. Scheduling policies for minimizing age of information in broadcast wireless networks[J]. IEEE/ACM Transactions on Networking, 2018, 26(6): 2637-2650.

[122] Jiang Z, Krishnamachari B, Zhou S, et al. Can decentralized status update achieve universally near-optimal age-of-information in wireless multiaccess channels?[C]. IEEE International Teletraffic Congress, 2018, 1: 144-152.

[123] Kalør A E, Popovski P. Timely monitoring of dynamic sources with observations from multiple wireless sensors[J]. arXiv preprint arXiv:2012.12179, 2020.

[124] Wang J, Liu Y, Das S K. Energy-efficient data gathering in wireless sensor networks with asynchronous sampling[J]. ACM Transactions on Sensor Networks (TOSN), 2010, 6(3): 1-37.

[125] Atay E U, Kadota I, Modiano E. Aging bandits: Regret analysis and order-optimal learning algorithm for wireless networks with stochastic arrivals[J]. arXiv preprint arXiv:2012.08682, 2020.

[126] Sharma H, Jain R, Gupta A. An empirical relative value learning algorithm for non-parametric mdps with continuous state space[C]. IEEE European Control Conference, 2019: 1368-1373.

[127] Hauskrecht M, Kveton B. Linear program approximations for factored continuous-state Markov decision processes[J]. Advances in Neural Information Processing Systems, 2003, 16.

[128] Li L, Littman M L. Lazy approximation for solving continuous finite-horizon MDPs[C]. AAAI, 2005, 5: 1175-1180.

[129] Liu K, Zhao Q. Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access[J]. IEEE Transactions on Information Theory, 2010, 56(11): 5547-5567.

[130] 3GPP Radio Access Network Working Group. Study on channel model for frequencies from 0.5 to 100 GHz (Release 15)[R]. 3GPP TR 38.901, 2018.

[131] Li Y, Zhang Y, Zhou H, et al. To relay or not to relay: Open distance and optimal deployment for linear underwater acoustic networks[J]. IEEE Transactions on Communications, 2018, 66(9): 3797-3808.

[132] Kam C, Kompella S, Nguyen G D, et al. Frequency selection and relay placement for energy efficiency in underwater acoustic networks[J]. IEEE Journal of Oceanic Engineering, 2013, 39(2): 331-342.

[133] Koseoglu M, Karasan E, Chen L. Cross-layer energy minimization for underwater ALOHA networks[J]. IEEE Systems Journal, 2015, 11(2): 551-561.

[134] Radosevic A, Ahmed R, Duman T M, et al. Adaptive OFDM modulation for underwater acoustic communications: Design considerations and experimental results[J]. IEEE Journal of Oceanic Engineering, 2013, 39(2): 357-370.

[135] Wan L, Zhou H, Xu X, et al. Adaptive modulation and coding for underwater acoustic OFDM[J]. IEEE Journal of Oceanic Engineering, 2014, 40(2): 327-336.

[136] Huang J, Diamant R. Adaptive modulation for long-range underwater acoustic communication[J]. IEEE Transactions on Wireless Communications, 2020, 19(10): 6844-6857.

[137] Xiaohong S, Haiyan W, Yuzhi Z, et al. Adaptive technique for underwater acoustic communication[M]. IntechOpen, 2012.

[138] Zhao H, Li X, Han S, et al. Adaptive relay selection strategy in underwater acoustic cooperative networks: a hierarchical adversarial bandit learning approach[J]. IEEE Transactions on Mobile Computing, 2021.

[139] Qarabaqi P, Stojanovic M. Statistical characterization and computationally efficient modeling of a class of underwater acoustic communication channels[J]. IEEE Journal of Oceanic Engineering, 2013, 38(4): 701-717.

[140] Brekhovskikh L M, Lysanov Y P, Beyer R T. Fundamentals of ocean acoustics[J]. 1991.

[141] Stojanovic M. On the relationship between capacity and distance in an underwater acoustic communication channel[J]. ACM SIGMOBILE Mobile Computing and Communications Review, 2007, 11(4): 34-43.

[142] Besson L, Kaufmann E. The generalized likelihood ratio test meets KL-UCB: an improved algorithm for piece-wise non-stationary bandits[J]. Proceedings of Machine Learning Research vol XX, 2019, 1: 35.

[143] Jiang L, Walrand J. A distributed CSMA algorithm for throughput and utility maximization in wireless networks[J]. IEEE/ACM Transactions on Networking, 2009, 18(3): 960-972.

致 谢

“雄关漫道真如铁，而今迈步从头越。”此时此刻，用这句话来形容博士四年的生活和感受再合适不过了。“春去秋来，花开花落。”无论如何，又到了和一段旅程说再见的时候。在这之前，感谢一路走来，指导过我、帮助过我、陪伴过我的每一个人。

首先，感谢我的导师付立群教授。感谢她在我博士求学路上对我进行的悉心指导和心理沟通。她带我一步步认识到如何独立思考、如何科研有方、如何逻辑写作，在我工作方式不足时，她总是耐心的鼓励我、引导我。她的科研态度和专注深深的影响了我。同时，感谢实验室其他老师：胡晓毅老师、王德清老师、吕江滨老师、游理钊老师、岳蕾老师和解永军老师在组会上给予的建议，以及生活上的帮助。

非常感谢美国休斯顿大学的韩竹教授。感谢他四年来对我的一路指导和关心，他的风趣、幽默、睿智为我提供了更加轻松愉快的科研环境。韩老师引导我面对陌生的科研问题要学会分解问题，鼓励我要敢于接受挑战，树立正确的科研态度。同时，感谢休大访学期间遇到的闵明慧博士，你在生活中的活力和对科研的热爱深深地感染了我；感谢张泓亮博士每次组会讨论时给出的宝贵意见；感谢和善友爱的 Jim Yuvancic 在生活上和科研上对我的极大帮助。感谢访学期间遇到的每一位伙伴：贾子晔、凌壮、杨杨、陈亚丽、陈大炜、杜勋胜、曹雪琳、庄子瑞、丁家豪、张心悦、张龙老师、孙恩昌老师等。

感谢实验室可爱又有爱的小伙伴们：刘圣波、钟凯琪、叶小文、黄嘉杰、王艺哲、余丹、李海宇、徐景鑫、王凤宇、黄愉芳、邱瑾、李徐竹、邱嘉航、全旭、张一帆、林童童、叶芬、汪梦雅、朱泽文、凌碧烽、吴占、韩兴隆、应忠翔、陈臻臻、赵信博、周倩、陈旭、秦娴、汤智榕、曹荣幽、李琛艳等，在我遇到科研问题的时候感谢你们帮我一起分析、讨论，感谢你们营造的快乐轻松的科研环境，陪我度过了难忘的科研时光。

最后，要特别感谢我的家人，感谢我的父母和姐姐对我默默的支持，感谢你们默默的付出不求回报。遇到困难和挫折时，家人的鼓舞和理解让我拥有了坚定的毅力，感谢你们始终如一的陪伴。

攻读博士学位期间取得的科研成果

1. 攻读博士期间发表的期刊和会议论文

[1] **Jingwen Tong**, Hongliang Zhang, Liqun Fu, Amir Leshem, and Zhu Han. Two-Stage Resource Allocation in Reconfigurable Intelligent Surface Assisted Hybrid Networks via Multi-Player Bandits [J], *IEEE Transactions on Communications*, 2022, May, 70(5), 3526-3541. **(JCR 2 区)**

[2] **Jingwen Tong**, Liqun Fu, and Zhu Han. Throughput Enhancement of Full-Duplex CSMA Networks Using Multi-Player Bandits [J], *IEEE Internet of Things Journal*, 2021, Mar. 3, 15(8):11807-11821. **(JCR 1 区)**

[3] **Jingwen Tong**, Liqun Fu, and Zhu Han. Age-of-Information Oriented Scheduling for Multi-Channel IoT Systems with Correlated Sources [J], *IEEE Transactions on Wireless Communications*, DIO: 10.1109/TWC.2022.3179305. **(JCR 1 区)**

[4] **Jingwen Tong**, Yizhe Wang, Liqun Fu, and Zhu Han. Model-Based Thompson Sampling for Frequency and Rate Selection in Underwater Acoustic Communications [J], *IEEE Transactions on Wireless Communications*, Submitted. **(JCR 1 区)**

[5] **Jingwen Tong**, Liqun Fu, and Zhu Han. Throughput Enhancement of Full-Duplex CSMA Networks via Adversarial Multi-Player Multi-Armed Bandit [C], *IEEE Global Communications Conference* (GLOBECOM), 2019, Dec. 9, Waikoloa, HI, USA. **(CCF C 类, EI)**

[6] **Jingwen Tong**, Shuyue Lai, Liqun Fu, and Zhu Han. Optimal Frequency and Rate Selection Using Unimodal Objective Based Thompson Sampling Algorithm [C], *IEEE International Conference on Communications* (ICC), 2020, Jun. 7, Dublin, Ireland. **(CCF C 类, EI)**

[7] Zhenzhen Chen, **Jingwen Tong**, Liqun Fu, and Zhu Han. Over-the-Air Computing Aided Federated Learning and Analytics via Belief Propagation Based Stochastic Bandits [C], *IEEE International Conference on Communications* (ICC), 2022, Accepted. **(CCF C 类, EI)**

2. 已授权的国家发明专利

[1] 付立群，**童景文**，岳蕾；一种利用 MAB 提升全双工 CSMA 网络吞吐量的方法；2019.06.14；CN110233762B。

[2] 岳蕾，赖舒悦，付立群，**童景文**；一种提升水声通信链路平均吞吐量的方法；2020.04.10；CN111431628B。

3. 参与的项目

[1] 基于非凸优化的无线网络能量效率与资源联合分配的研究（No. 61771017），国家自然科学基金（面上）项目，2018-2021，参与。

[2] 福建省组织部, 纵向项目，福建省百人计划科研项目，2019-2021，参与。